(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 January 2006 (05.01.2006)

PCT

(10) International Publication Number
WO 2006/000410 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/EP2005/006795

(22) International Filing Date: 23 June 2005 (23.06.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PCT/EP2004/006781 23 June 2004 (23.06.2004) EP

(71) Applicants *(for all designated States except US)*: **BAYER BIOSCIENCE N.V.** [BE/BE]; Technologiepark 38, B-9052 Gent (BE). **BAYER CROPSCIENCE S.A.** [FR/FR]; 16 rue Jean-Marie Leclair, F-69009 Lyon (FR). **COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANISATION** [AU/AU]; Limestone Avenue, Cambell, ACT 2601 (AU).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **BOUROT, Stéphane** [FR/FR]; 16 rue Robert Desnos, F-77370 Nangis (FR). **GIGONZAC, Olivier** [FR/FR]; 114 ter rue de Silly, F-92100 Boulogne-Billancourt (FR). **METZLAFF, Michael** [DE/BE]; Irislaan 26, B-3080 Tervuren (BE). **WATERHOUSE, Peter** [AU/AU]; 5 Banjine Street, O'Connor, Canberra, ACT 2602 (AU). **HELLIWELL, Christopher** [AU/AU]; 167 Wattle Street, O'Connor, Canberra, ACT 2602 (AU).

(74) Agents: **BRANTS, Johan, Philippe, Emile** et al.; De Clercq, Brants & Partners CV, E. Gevaertdreef 10A, B-9830 Sint-Martens-Latem (BE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:
— *of inventorship (Rule 4.17(iv)) for US only*

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHOD FOR IDENTIFYING NUCLEIC CAPABLE OF MODULATING A GENE IN A BIOLOGICAL SYSTEM

(57) Abstract: The present invention relates to a method for identifying one or more regions of a gene of interest which are suitable targets for modulating in a biological system of interest. Said regions enable specific targeting of a gene, or of a family of genes, while minimising any effects on other genes in the biological system. The invention further relates to a computer program and a kit therefor.

METHOD FOR IDENTIFYING NUCLEIC CAPABLE OF MODULATING A GENE IN A BIOLOGICAL SYSTEM

BACKGROUND TO THE INVENTION

Post-transcriptional gene silencing (PTGS) in plants and animals is achieved by introducing into cells double stranded RNA (dsRNA) homologous to the gene of interest. Silencing occurs by the selective targeting and degradation of the homologous mRNA. The RNA interference (RNAi) can achieve selective hydrolysis of cellular mRNA, even when only sub-stoichiometric amounts of double stranded RNA are present. In some systems, RNAi can maintain selective gene silencing, even when there is a 50-to 100-fold increase in cell mass.

A model for the mechanism of RNAi is based upon the observation that the introduced dsRNA is bound and cleaved by RNAseIII-like endonuclease (DICER) to generate 21 and 22-nucleotide double stranded RNA products. These small interfering RNAs (siRNAs) remain complexed with the endonuclease. The resulting dsRNA-protein complexes appear to effect the selective degradation of homologous mRNA. It has been established that duplexes of 21-nucleotide RNAs are sufficient to suppress expression of endogenous genes in mammalian cells. This was demonstrated by selective silencing of endogenous lamin A/C expression in human epithelial cells following introduction of the cognate siRNA duplex. Thus, introduction of siRNA into mammalian cells is sufficient to selectively target homologous mRNA and silence gene expression.

RNA for silencing may be introduced into a organism as a DNA clone, coded as a stem-loop in which the stem comprises dsRNA. When the clone is transcribed, the stem-loop so formed targets homologous mRNA and silences gene expression. For efficient silencing, the length of the stem may be 300 nucleotides pairs in length or longer; in such cases the stem-loop is processed by the organism, for example by DICER or DICER-like RNAses into 21nt siRNAs, which effect gene silencing as mentioned above. The latter process is sometimes referred to as DNA directed RNA interference or ddRNAi.

There are presently no methods for determining the sequences of double stranded RNA suitable for PTGS. The prior art (*e.g.* WO 03/70966) describes a method for selecting double stranded RNA sequences wherein a siRNA library is used. However, such a method requires extensive experimentation and screening to arrive at a suitable inhibitory double stranded RNA. Such technique can be laborious and costly.

Other methods of the prior art relate to the silencing of specified genes (*e.g.* WO 03/066650, WO 03/070983, WO 03/070969, WO 03/070968 and WO 03/070918) and do not disclose a generally applicable method for selecting double stranded RNA suitable for PTGS.

There is presently no method to test prior to experimentation, which regions of a gene of interest would be suitable for targeting by double stranded RNA.

Furthermore, there are no methods for determining which regions of a gene of interest would be suitable for targeting by ddRNAi.

Furthermore, there is presently no method in the art to assess the impact of an inhibitory double stranded RNA upon other genes in the genome of a cell. Following the methods of the art, the impact of such double stranded RNA on other targets would require experimentation, which is technically demanding and costly.

Furthermore, there is presently no method in the art for designing double stranded RNA suitable for PTGS that can selectively target a family of genes, without targeting other genes in the genome. For example, a family of enzymes, control proteins etc.

Furthermore, there is presently no method in the art for designing DNA suitable for cloning into a vector, said vector suitable for PTGS that can selectively target a specific gene or family of genes, without targeting other genes in the genome.

Therefore, there is a need for a method which indicates the regions of a gene of interest that are suitable for targeting by a nucleic acid suitable for PTGS. Furthermore, there is a need for a method which avoids the targeting other genes in the genome.

SUMMARY OF THE INVENTION

One embodiment of the present invention is a method for identifying nucleic acid capable of modulating a specific gene of interest, GOI in a biological system of interest, SOI, comprising:

(a) obtaining a sequence of the GOI,

(b) identifying genes in the SOI that share a nucleotide identity with the GOI,

(c) determining one or more regions of the GOI that have a divergent identity with the genes identified in step (b), and

(d) identifying, from the regions determined in step (c), nucleic acid capable of modulating a specific GOI in a SOI.

Another embodiment of the present invention is a method as described above wherein a gene of said SOI of step (b) is identified by determining whether any window of at least 12 nucleotides of one strand of the GOI, exhibits identity to said gene of the SOI.

Another embodiment of the present invention is a method as described above wherein step (c) further comprises the following steps:

(c1) determining which windows of at least 12 nucleotides in length of both strands of the GOI exhibit identity with any of the genes of the identified in step (b), said identity allowing at least one mismatch, and

(c2) providing regions of the GOI devoid of said nucleotide windows identified in step (c1).

Another embodiment of the present invention is a method as described above wherein the identity in step (b) is between 17 and 25 nucleotides.

Another embodiment of the present invention is a method as described above wherein the identity in step (c1) is between 17 and 25 nucleotides, and the number of mismatches is 1.

Another embodiment of the present invention is a method as described above, wherein the step (b) comprises the use of a statistical evaluation of homology.

Another embodiment of the present invention is a method as described above, wherein the step (b) comprises the use of the "BLAST" algorithm or variant thereof.

Another embodiment of the present invention is a method as described above, wherein the step (c) comprises the use of the "COMPARE" algorithm from the GCG package, or a variant thereof.

Another embodiment of the present invention is a method as described above, wherein step (c1) treats sequences which are non-identical, but can hybridise to the same oligonucleotide *via* G:U or G:T mismatching as exhibiting identity.

Another embodiment of the present invention is a method as described above, wherein said non-identical sequences are identified using the "NUCFUZZ" algorithm from the EMBOSS package, or a variant thereof.

Another embodiment of the present invention is a method as described above, wherein the best region identified in step (d) is submitted to the method as described above, so the G/T rule is implemented after the best sequence has been found.

Another embodiment of the present invention is a method as described above further comprising a step of concatenating two or more regions of the GOI determined in step (c) when the maximum length of a region determined in step (c) is less than a minimum number of nucleotides determined by the user.

Another embodiment of the present invention is a method as described above further comprising a step of determining the sequence of at least one duplex RNA, wherein one strand of said RNA duplex corresponds to at least part of a divergent region determined in step (c).

Another embodiment of the present invention is a method as described above comprising the step of determining potential secondary structure-forming regions of said duplex RNA.

Another embodiment of the present invention is a method as described above further comprising a step of determining at least one DNA sequence suitable for cloning into a vector, wherein said DNA sequence corresponds to at least part of a divergent region determined in step (c).

Another embodiment of the present invention is a method as described above, wherein said vector is capable of producing double stranded RNA in the SOI.

Another embodiment of the present invention is a method as described above, wherein said RNA duplex comprises at one substitution where U is C, C is U, G is A, or A is G.

Another embodiment of the present invention is a method as described above further comprising a step of determining the sequences of at least one pair of PCR primers, suitable for amplifying the DNA sequence(s) as described above.

Another embodiment of the present invention is a method as described above, wherein

- steps c1) to d) are repeated, and
- the number of mismatched permitted in step c1) is increased by one after each repeat,

so producing a list regions of the GOI that have a divergent identity with the genes identified in step (b), corresponding to an increasing number of permitted mismatched.

Another embodiment of the present invention is a method for identifying nucleic acid capable of modulating a family of genes in a biological system of interest, SOI, comprising the steps of:

(A) obtaining the sequences of the genes in the family of genes of interest, FGOI,

(B) calculating a single homologous sequence from the sequences of step (A),

(C) selecting each region of homology calculated in step (B) as a GOI,

(D) identifying nucleic acid capable of modulating the FGOI by using each GOI of step (C) in a method according as described above.

Another embodiment of the present invention is a method as described above, wherein the regions of homology in step (C) are concatenated to form a single GOI.

Another embodiment of the present invention is a method as described above, wherein a single homologous sequence of step (B) is calculated using Clustal W.

Another embodiment of the present invention is a method as described above, wherein a single homologous sequence of step (B) is a best representative sequence, BRG, calculated by comparing genes of the FGOI a pair at a time.

Another embodiment of the present invention is a method as described above, wherein the BRG is calculated by:

(I) selecting a sequence of 18 to 20 nt from the start of a first gene of the FGOI,

(II) comparing said sequence across each of the other genes of the FGOI for an identity match, or no identity match,

(III) repeating step (I) using the next gene of the FGOI, until all the genes of FGOI have been exhausted, and

(IV) selecting the gene with the highest number of identity matches and the lowest number of no identity matches, which is the BRG .

Another embodiment of the present invention is a computer program stored on a computer readable medium, capable of performing a method as described above.

Another embodiment of the present invention is a computer program as described above comprising an ability to display a user interface on a computer, said interface allowing a user to provide one or more parameters to a method of the invention.

Another embodiment of the present invention is a computer program as described above further comprising an ability to display a user interface on a computer, said interface allowing a user to receive an indication of the sequences provided by a method of the invention.

Another embodiment of the present invention is a computer program as described above further comprising an ability to make the sequences provided by a method of the invention available by email, by Internet publishing, or by storing on a networked computer.

Another embodiment of the present invention is a computer program as described above further comprising one or more databases of gene sequences of the SOI.

Another embodiment of the present invention is a computer program as described above, further comprising an ability to display said user interface on a web browser of a remote computer connected to the Internet.

Another embodiment of the present invention is a computer readable storage device on which a computer program as described above is stored.

Another embodiment of the present invention is a kit comprising at least one computer readable storage device as described above and one or more vectors suitable for use in gene modulation.

Another embodiment of the present invention is a kit as described above wherein said vectors are any capable of producing double stranded RNA in the SOI.

Another embodiment of the present invention is a kit as described above wherein said vectors are one or more of pDON, pHellsgate, pHellsgate 8, pHellsgate 11, and pHellsgate 12.

Another embodiment of the present invention is unknown nucleic acid identified or determined according to a method as described above.

SUMMARY OF THE FIGURES

Figure 1: A flow chart indicating an example of a method of the invention for performing step (b) described below.

Figure 2: A flow chart indicating an example of a method of the invention for performing part of step (c) described below.

Figure 3: A flow chart indicating an example of a method of the invention for performing part of steps (c) and (d) described below.

Figure 4: A flow chart indicating another example of a method of the invention for performing step (b) described below.

Figure 5: A flow chart indicating another example of a method of the invention for performing part of step (c) described below.

Figure 6: A flow chart indicating another example of a method of the invention for performing part of steps (c) and (d) described below, including the FUZZNUC algorithm.

Figure 7: A flow chart indicating another example of a method of the invention for performing part of steps (c) and (d) described below.

Figure 8: A flow chart indicating an example a method for providing the best RNA duplexes from the list of RODI obtained using a method of the invention.

Figure 9: A flow chart indicate an example of a method steps (a), (b) and (c) described below.

Figure 10: A flow chart indicate an example of a method steps (a), (b) and (c) described below, including the FUZZNUC algorithm.

Figure 11: A flow chart depicting an example of steps of a method according to the invention wherein either a single gene or a family of genes is to be modulated.

Figures 12 to 14: Flow charts depicting example of steps of a method according to the invention wherein a family of genes are to be modulated.

Figures 15 to 24: Examples of screen shots of a computer program of the invention capable of displaying a web page on a remote computer.

DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to a method for modulating the expression of a gene of interest (GOI) or a family of genes of interest (FGOI) wherein said modulation is specific for the GOI or FGOI. Said gene modulation may be post-transcriptional gene silencing.

The inventors of the present invention have discovered that a method which comprises a first step of examining the genome database of the species or system of interest (SOI) for genes containing an identity of 12 nucleotide or higher with the GOI (filtering step), and a second step of searching the filtered genes for an identity of at least 12 nucleotides and optionally at least 1 mismatch provides a rapid and accurate means for determining the regions of the GOI which have a divergent identity with other genes in the genome of the SOI.

By divergent identity in respect of a region of a GOI sequence means that the region is sufficiently different from any sequence of the SOI, such that no window of at least 12 nucleotides in length of both strands of the divergent region exhibits identity with the SOI. The size of the window may be 12, 13, 14, 15, 16, 17, 18, 19,. 20, 21, 22, 23, 24, 25, or more than 25 nucleotides, and preferably 18, 19, 20, 21, 22, or 23 nucleotides, more preferably 21 or 19 nucleotides. A window exhibiting identity may have an absolute sequence match, or a match allowing at least one mismatch, the number of mismatches being 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12 or more than 12, and preferably 1.

From the regions of the GOI which have a divergent identity with other genes in the genome of the SOI, one or more double stranded RNA duplexes are designed which modulate specifically the GOI.

One embodiment of the present invention relates to a method for identifying nucleic acid capable of modulating a gene in a SOI comprising the steps of:

(a) obtaining a sequence of the GOI,

(b) identifying genes in a SOI that share at least a 12 nucleotide identity with the GOI,

(c) determining the regions of the GOI that have a divergent identity with the genes identified in step (b).

(d) optionally designing at least one duplex RNA capable of binding to a region of divergent identity determined in step (c).

The inventors have found that the method allows a reliable and fast identification of targetable regions of the specific GOI, so enabling the design of at least one duplex RNA that is capable of modulating the activity of the specific gene. The invention reduces the impact of the duplex RNA so designed upon modulation of other genes in the system.

The SOI may be any biological system capable of exhibiting post transcriptional RNA-induced gene silencing. According to one aspect of the invention, the SOI may be an organism that comprises the GOI. According to another aspect of the invention, the SOI may be a combination of two or more organisms, one or more of which comprises the GOI, and one or more of which is naturally devoid of the GOI; the combination may be temporary or permanent. According to one aspect of the invention, the SOI may be an organism that is naturally devoid of the GOI. Where the SOI is, for example, a combination of two or more organisms, an insect GOI may be silenced by administration of a gene silencing compound to a plant on which it feeds. Preferably, the SOI has been sequenced such that the sequences of all known cDNAs therein are available in a computer-readable format such as, for example, on an electronic database and the below-mentioned filtering step may be applied against this subset of sequences of the SOI. Examples of SOIs include, but are not limited to animals (*e.g.* human, mouse, rat, horse, pig, cattle, chicken, nematodes, drospohila), plants (*e.g.* arabidopsis, rice, cotton, canalo, tobacco), yeasts, moulds and fungi, cells and tissues thereof.

The list of SOI includes present and future-available transcriptomes. Therefore, included are SOI for which the sequence of the complete genome is not yet available. As used herein, "a transcriptome" refers to a set of nucleotide sequences selected from SOI which correspond to the transcribed regions of the genome of the SOI. A transcriptome may correspond to the nucleotide sequences of the clones of a cDNA library of the SOI. The more complete the transcriptome, the more efficient the current method will allow to determine the regions of a GOI suitable for targeting by ddRNA.

Filtering step

According to a method of the invention, genes of an SOI which share a region of at least 12 nucleotides identity with the GOI are identified in step (b) (Figure 1). The length of sequence identity may be 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more than 25 nucleotides, and preferably 21 nucleotides. Genes of the SOI which share an identity with the GOI may be calculated using known methods such as, for example, comprising the use of BLAST (Basic Local Alignment Search Tool , Altschul et al. 1990, *J. Mol Biol.* 215: 403-410), a variation of BLAST, or similar software packages available from the GCG Wisconsin package (see for example http://www.accelrys.com/products/gcg_wisconsin_package), or from DNAStar (http://www.dnastar.com/).

Preferably, the genes identified in step (b) are identified comprising the use of BLAST or a variation thereof, followed by one or more parsing routines to discard sequences exhibiting less than 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 or 25 nucleotides identity with the GOI (Figure 4). When using BLAST (blastn), the wordsize is preferably set to eleven, and the threshold value (EXPECT) is ten. Most preferably, the BLAST parameters are set at default, with the exception that the threshold value '–e' is set to 150, –b = 5000 (Number of database sequence to show alignments), –v = 5000 (Number of database sequences to show one-line descriptions), and –F =F (Filter query sequence is out). Variations and optimisations as known to the skilled person are within the scope of the present invention.

According to one aspect of the invention, the step (b) is performed by determining the regions of perfect matching (*i.e.* absolute identity), and the regions of imperfect matching (*i.e.* with mismatches, insertions and/or deletions) between the GOI and a gene of the SOI. The results for both types of matching are analysed to determine whether the GOI and a gene of the SOI are statistically similar, compared with the result expected by a random match. Once the genes of the SOI have been compared, those statistically similar to the GOI are further filtered to discard any sequences which contain an identity of less than 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 or 25 nucleotides identity with the GOI.

According to the invention, filtering step (b) may be performed using any algorithm which performs a statistical analysis of local homology searches. Further details of BLAST may be found in, for example, The Statistics of Similarity Scores available at http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html#head2 which provides an overview of the underlying theory. Algorithms similar to BLAST include, for example, WuBLAST (Washington University), FASTA (Pearson, W. R. (1999) Flexible sequence similarity searching with the FASTA3 program package. Methods in Molecular Biology; W. R. Pearson and D. J. Lipman (1988), Improved Tools for Biological Sequence Analysis, PNAS 85:2444-2448; W. R. Pearson (1998) Empirical statistical estimates for sequence similarity searches. In J. Mol. Biol. 276:71-84; and Pearson, W. R. (1996) Effective protein sequence comparison. In Meth. Enz., R. F. Doolittle, ed. (San Diego: Academic Press) 266:227-258.)

The inventors have found that such a statistical analysis allows genes to be filtered much more rapidly than using a non-statistical approach.

According to one aspect of the invention, the coding strand (+) of the GOI is compared with the both coding (+) and non-coding (-) strands of genes in the SOI.

When two sequences share a region of identity, it means both sequences share a nucleotide region identical in sequence and length. *E.g.* ATG ATG ATG ATG CC and ATG ATG ATG GG share a 9 nucleotide identity. Unless otherwise stated, identity is without mismatches. When a degree of identity is permitted, it may mean that one or more mismatches are allowed *e.g.* ATG ATC ATG ATG CC and ATG ATG ATG GG share a 8 nucleotide identity with 1 nucleotide mismatch.

When a degree of identity is permitted, it may also mean that one or more deletions are allowed *e.g.* ATG AT_ ATG ATG CC and ATG ATG ATG GG share a 8 nucleotide identity with 1 nucleotide deletion.

When a degree of identity is permitted, it may also mean that one or more substitutions are allowed *e.g.* ATG ATGG ATG ATG CC and ATG ATG ATG GG share a 8 nucleotide identity with 1 nucleotide substitution.

Genes which belong to the same genome as the GOI may be available in an electronic database. For example, the human genome is available electronically as is the genome of animals such as horse, pig, mouse, cattle, and chicken. The genome of *Arabidopsis* is available electronically as well as the sequence of the transcriptome thereof (*e.g.* from the *TAIR_At_transcript* database). Preferably, an electronic database of an SOI contains the sequences for all the genes of the SOI. Thus, it is an aspect of the invention that step (b) comprises a query of an electronic database containing the GOI. It is another aspect of the invention that the GOI or the FGOI are excluded as members of the SOI when performing a method of the invention.

Compare step

Another embodiment of the present invention is a method as described herein, wherein the regions of divergent identity determined in step (c) are calculated by searching the sequence of each gene identified in step (b) for identity with the GOI (Figure 2). Those regions of the GOI which bear little or no identity to the genes identified in step (b) are considered regions of divergent identity (RODI).

According to an embodiment of the present invention the comparison may be performed by any method which determines which windows of at least 12 nucleotides in length of both strands of the GOI exhibit identity with any of the genes of the identified in step (b), said identity permitting at least one mismatch.

The degree of identity may be adjusted by the user *i.e.* the stringency may be altered to account for mismatches, deletions and/or insertions.

The size of the window in step (c) may be 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or greater than 30 nucleotides, preferably 18, 19, or 20 nucleotides, but most preferably 19 nucleotides.

The number of non-matching nucleotides permitted may be 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, or 29, preferably 1, 2 or 3 nucleotides and most preferably 1 nucleotide. Alternatively, the number of permitted mismatches is less than 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9% or 10% of the length of the window length. It is an option of the invention that mismatches which comply with the G/T rule are ignored (see 'Allowable mismatches' below).

The number of deletions permitted may be 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, or 29, preferably 0, 1, 2 or 3 nucleotides and most preferably 0 nucleotides. Alternatively, the number of permitted deletions may be less than 0%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9% or 10% of the length of the window.

The number of insertions permitted may be 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, or 29, preferably 0, 1, 2 or 3 nucleotides and most preferably 0 nucleotides. Alternatively, the number of permitted insertions may be less than 0%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9% or 10% of the length of the window.

The window may be moved by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 nucleotides, and preferably 1 nucleotide after each comparison, or after comparing each of the genes identified in step (b).

According to one aspect of the invention, both the coding (+) and anti-sense (-) strands of the GOI, and the coding strand (+) of the genes identified in step (b) are used in step (c).

According to one aspect of the invention, step (c) further comprises the following steps:

(c1) determining which windows of at least 12 nucleotides in length of both strands of the GOI exhibit identity with any of the genes of the identified in step (b), said identity permitting at least one mismatch, and

(c2) providing regions of the GOI devoid of said nucleotide windows identified in step (c1).

Once the GOI windows bearing full or near-identity to the genes identified in step (c) have been obtained, these may be used to obtain regions of identity of the GOI (windows which showed identity) and regions of the GOI with divergent identity (windows which showed no/little identity). Thus, RODI are the remains of the GOI once the regions of identity have been removed.

Where both the coding (+) and non-coding strands (-) of the GOI are searched, regions of divergent identity are determined as being free from identity on both the (+) strand and the corresponding (-) strand.

Preferably, the regions of identity are determined comprising the use of the "Compare" algorithm from the GCG package, or a variation thereof (Figure 5). The inventors have found that by enforcing a WORdsize of at least 15, and a degree of stringency which permits some non-complementary base pairs to match, the regions determined by the method are suitable for design of RNA duplex that is stringent for the gene of interest. According to one embodiment of the invention, the WORdsize may be between 12 and 25, and is preferably 18, 19, or 20, and is most preferably 19. The STRIngency may be between 15 and 25 and is preferably 17, 18, 19 and most preferably 18.

Preferably, the output of the "Compare" algorithm is parsed, and the regions of divergent identity are identified from the GOI in a localization step.

In another aspect of the invention step (c) may be performed by any method which is capable of searching for identity, with a given stringency within a sequence. Alternative methods include routines in software packages such as those available from , for example, DNAStar (http://www.dnastar.com/) or EMBOSS http://www.hgmp.mrc.ac.uk/software/EMBOSS/apps).

Allowable mismatches

A mismatched base pair, wherein non-Watson-Crick base pairing still forms a stable duplex, may be an allowable mismatch according to one aspect of the invention *i.e.* such mismatch is ignored. For example, according to the G/T rule, the base pairs G:U and G:T form stable base pairs owing to the orientation of the bases and hydrogen bond donors and acceptors. Therefore, a region between the GOI and a gene of an SOI which is non-identical, but can hybridise to the same oligonucleotide *via* G:U or G:T mismatching may be considered as having perfect identity according to one aspect of the invention. For example, the sequences ATG ATT ATG ATG CC and ATG ATC ATG GG may be considered identical, with no mismatches for the reason that a therapeutic RNA, 3'-UAC UAG UAC UAC GG-5' can bind to both in the RNA form, *via* G:U mismatching. Similarly, the sequences ATG ATG ATG ATG CC and ATG ATA ATG GG may be considered identical, with no mismatches for the reason that therapeutic RNA of sequence 3'-UAC UAU UAC UAC GG-5' can bind to both in the RNA form, *via* G:U mismatching.

One aspect of the present invention is a method as described, wherein step (c) searches the sequence of each gene identified in step (b) for identity with the GOI, where i) T is read as T or C, ii) C is read as C or T, iii) G is read G or A, or iv) A is read as A or G in either the GOI or a gene of the SOI. Alternatively, step (c) may proceed with any two or more of i) to iv). Preferable, step (c) includes the step where T is read as T or C in either the GOI or a gene of the SOI. It is an aspect of the invention that the size of the window used in step (c) is 21 nt where the G/T applied. By checking for of G:T mismatches, the possibility of cross-reactivity is further reduced.

Allowable mismatches as described above may be accounted in step (c) by using the "FUZZNUC" algorithm from the EMBOSS package. FUZZNUC performs the task of determining regions of identity with a definable window size, number of mismatches, deletions, insertions, and accounting for allowable mismatches such as defined in the G/T rule above. The "Compare" algorithm (GCG) mentioned earlier also performs the task of determining regions of identity with a definable window size, number of mismatches. Therefore, when the FUZZNUC algorithm is used, the "Compare" algorithm may be redundant. An example of an implementation of the present method using FUZZNUC is given in Figure 6.

FUZZNUC may be incorporated into an algorithm (COMPFUZ) which selects a short domain within a GOI for a FUZZNUC calculation against the SOIs, and which incrementally changes the position of said domain. COMPFUZ is described in Figures 6 (**65** to **68**), 10 (**1011**, **1015**), 13B (**1321**) and 14 (**144** to **147**).

Implementing FUZZNUC or COMPUZ may be computationally expensive for some computers which implement the invention. Another aspect of the present invention, therefore, permits a user to find the best RODI where the G/T rule is not applied, for example using 'Compare', and then submitting the best RODI, which is shorter than the GOI, to the method of a present invention where the G/T rule is applied, for example, using FUZZNUC. In cases where the calculations involving the G/T rule are expected to be lengthy, the invention may provide the option to provide the results as a written file or by email, or by any means of making results available. Where the invention is implemented in a workstation environment or via a web page, for example, the user can have the option to set up a calculation overnight and go offline or close the browser window, leaving the method running. The results are viewed by subsequently inspecting the written file or email message for instance. An example of an interface which presents the email option is given in Figure 23.

Regions of divergent identity

The invention can present the regions of divergent identity in various ways (Figure 7). In one embodiment of the invention, the regions of divergent identity are ranked according to length *e.g.* from longest, uninterrupted divergent region to shortest.

In another embodiment of the invention, the GOI is presented so that the regions of divergent identity are indicated thereon, *e.g.* as lower case lettering, non-ATCG lettering, underlined nucleotides, etc.

In another embodiment of the invention, the GOI is presented so that the regions of identity are indicated thereon, *e.g.* as lower case lettering, non-ATCG lettering, underlined nucleotides, etc.

In another embodiment of the invention, a RODI is presented which is the most unique region of the SOI and having the most mismatches. According to another aspect of the invention, a list of RODI is ranked according to the number of mismatches permitted in step c) or c1). The RODI corresponding to 0 mismatches, then 1 mismatch, then 2 mismatches, etc in step c) or c1), are displayed, the number of mismatches increasing until no more RODI can be found, or earlier. Such mode is known as 'TouchDown' in view of the incremental number of permitted mismatches.

In another embodiment of the invention, the GOI is presented so that the regions of homology and divergent identity indicated thereon, *e.g.* as coloured bars below. The best sequence may be indicated in one colour, regions of homology in another, and RODIs in another. Genes identified in step (b) above may by indicated in a schematic alignment below which indicates regions of homology and optionally RODIs.

In one embodiment, a method of the invention may indicate a region of divergent identity of a GOI as a combination of two or more regions when the length of the longest divergent region falls below a minimum length (Figure 8). For example, a user may specify that he wishes to find a minimum length of a divergent region of a GOI of 300 nucleotides. However, in practice, a GOI may, for example, have a maximum of 250 nucleotides of uninterrupted divergence from the genes of the SOI. To provide the closest solution, the method may combine this 250 nucleotide region with another divergent region of more than 49 nucleotides, if available, to provide a combined divergent region of 300 nucleotides or greater.

Once in possession of the regions of divergent identity, one or more an RNA duplexes may be designed that are capable of binding to said region.

Family of genes

Another embodiment of the invention is a method as described herein, for identifying duplex RNA suitable for modulating a family of genes of interest (FGOI). According to this aspect of the invention, regions of homology between the members of the FGOI are determined. A single sequence is formed which bears the closest homology to all the members of the FGOI. Homology may be calculated by known methods (Figure 11).

According to one aspect of the invention, homology is calculated using a system of matrix built from an exhaustive comparison of the genes of the FGOI, two by two. This is elaborated in Figures 13A, 13B and 14, which single sequence is a best representative gene (BRG) chosen from the genes of the family. When genes of the FGOI are compared two by two, a sequence of 15, 16, 17, 18, 19, 20, or 21 nt, and preferably 19 nt is selected from the start of a first (n) gene of the FGOI, and it is compared across each of the other (m) genes of the FGOI. If there is an identity match, it is indicated, preferably in a matrix (common domains matrix, CD^{nm}). Another matrix preferably indicates absence of identity (Without hit matrix, WO^{nm}). Once the other genes have been scanned, the position of the sequence in the first gene is incremented by one nucleotide, and identity against the other genes assessed again. Once the window has been moved across the first gene, the process is repeated again with the second (n+1) gene of the FGOI. To avoid repetition and save computational time, the matrices formed may be 'half-diagonal' which can be achieved, for example, by comparing a gene of the FGOI with those only listed above and not below. For example, the third gene would be compared with the forth, fifth, sixth etc. gene, but not with the second or first. The resulting matrices, CD and WO may respectively indicate the number of 'common domains hits' and 'no hits' for each pair of genes. The best representative gene is chosen from the gene which has the highest CD score, and the lowest WO score. Figure 24 shows the result of a two-by-two comparison.

According to another aspect of the invention, homology is calculated using a one or several multiple alignments, using algorithms such as ClustalW ("CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice" Julie D. Thompson, Desmond G. Higgins and Toby J. Gibson European Molecular Biology Laboratory, on http://bimas.dcrt.nih.gov/clustalw/clustalw.html) or T-coffee (T-Coffee: A novel method for multiple sequence alignments. C. Notredame, D. Higgins, J. Heringa, Journal of Molecular Biology,Vol 302, pp205-217,2000).

Those regions of the single sequence which show little/no homology to the FGOI are excised from the single sequence. The shortened, single homologous sequence so obtained is used in as the GOI in the method for identifying a region of divergent identity as described herein.

Alternatively, those regions of the single sequence which show an homology to the FGOI are each separately used in as the GOI in the method for identifying a region of divergent identity as described herein.

Alternatively, the single sequence is submitted as the GOI in the method for identifying a region of divergent identity as described herein.

Design of RNA duplex

In another embodiment of the invention, the method provides proposed designs for one or more RNA duplexes suitable for use as a modulator of a GOI. RNA duplexes may be designed using known methods and criteria. For example, one strand of the RNA duplex should be capable of binding to the coding strand of DNA of the GOI, and the other strand of the RNA duplex should be capable of binding to the antisense strand of the DNA of the GOI. Said binding is in a region of divergent identity, identified according to the invention.

The duplex RNA may comprise unmodified and/or modified nucleotides. The duplex RNA may be a stem-loop, or double stranded RNA. Optionally, the strands of the double stranded RNA may be tethered by chemical linkage at one or both ends. The duplex RNA may have a length sufficient to maintain a specific interaction.

According to one aspect of the invention, duplex RNA may have a length of 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, > 27, >100, >200, >300, >400, >400, >600, >700 or >800 base pairs.

According to one aspect of the invention, the designed RNA duplex may be as long as a region of divergent identity identified in a method of the invention. According to another aspect of the invention, the designed RNA duplex may be shorter than a region of divergent identity identified in a method of the invention. According to another aspect of the invention, the designed RNA duplex may be longer than a region of divergent identity identified in a method of the invention.
According to another aspect of the invention, the shortest RNA duplex may be designed to be 12 nucleotides in length. Preferable the length of RNA duplex is designed between 19 and 800 nucleotides.

The RNA duplex so designed may optionally be screened for possible RNA secondary structure-forming regions; this is effected using known methods. It is an aspect of the invention that an indication of secondary structure is provided by the methods and/or a ranking of suitable sequences wherein secondary structure-forming regions are accounted.

According to one aspect of the invention, the RNA duplex may comprise one or more mismatches according to the G/T rule *i.e.* i) U is U or C, ii) C is C or U, iii) G is G or A, or iv) A is A or G. Alternatively, RNA duplex may comprise at least one mismatch which is two or more of i) to iv). Preferable, a mismatch in the RNA duplex is one or more U, which may be substituted with C.

Where the method provides the combined sequence of two or more divergent regions to form a single divergent region of more than 300 nucleotides, an RNA duplex designed according to the invention may encompass the separate divergent regions.

In one embodiment, the duplex RNA designed according to the invention is suitable for introduction into the SOI as naked nucleic acid *i.e.* without additional residues.

DNA

In another embodiment, the duplex RNA designed according to the invention is suitable for introduction into the SOI by means of a nucleic acid vector such as a plasmid, a phage, a phagemid, a cosmid etc. To facilitate the ease of constructing a vector, the method may provide the sequence of nucleic acid, suitable for cloning into said vector. Optionally, the method may provide an indication of restriction sites, DNA linkers, primers for PCR reaction suitable to amplify the RODI etc to assist with the cloning.

To still assist the cloning of modulating polynucleotide into a vector, it is one aspect of the invention that the method presents one or more coding strands (+) of DNA, identified according to the invention, as an option for cloning into a vector. According to another aspect of the invention, the method presents one or more antisense strands (-) of DNA as an option for cloning into a vector. According to another aspect of the invention, the method presents both the coding (+) and antisense strand (-) as a single gene, said (-) and (+) strands tandemly arranged. According to another aspect of the invention, said (-) and (+) strands being tandemly arranged, may be separated by a linking sequence such as, for example, a hairpin loop or intron sequence. Where an intron sequence is used as a linker, efficiency of the silencing is further enhanced.

The inventors have found that when a 300nt region of divergent identity of the GOI is identified and cloned in a vector as a tandem arrangement of a (+) and (-) strand, the clone so formed leads to a generation of an RNA stem-loop in a SOI, wherein the stem is approximately 300 nucleotides in length. Said stem-loop RNA is processed by the SOI into short 21 nt siRNAs, so leading to efficient gene silencing.

Optionally, the tandemly arranged (+) and (-) strands in the DNA clone are operably linked to a promoter sequence in the vector.

Optionally, a DNA clone comprises one or more repeats of the above mentioned tandem arrangement of (+) and (-) strands.

It is a further aspect of the invention that the method proposes designs of PCR primers suitable for amplifying DNA for insertion into a vector, said DNA sequences based upon the divergent regions identified in step (c).

Computer program

Another embodiment of the present invention is a computer program stored on a computer readable medium capable of performing a method according to the invention. Another embodiment is a computer comprising a computer readable medium on which said computer program is stored.

Another embodiment of the present invention, is a computer program as described herein, further comprising the ability of displaying a user interface on a computer, said interface allowing a user to send an indication of the GOI to a method of the invention. Said indication of the GOI may be the sequence, or the name of the gene or other indication.

The user interface may also permit a user to indicate the SOI, or the database. The interface may also permit a user to indicate parameters for the method, such as, for example, stringency, wordsize, preferred ranges of divergent region to display, output format, parameters relating to secondary folding, and/or parameters relating to PCR primers.

Another aspect of the invention is a computer program as described above further comprising the ability to display a user interface a computer, said interface allowing a user to receive an indication of the divergent regions of the GOI determined in step (c) in appropriate formats, proposals for designs of RNA duplex, proposal for designs of DNA suitable for insertion into a vector, and/or proposal for the design of PCR primers. Other indications may be included which are derived from the information provided by the method and would assist with PTGS, such as, for example, regions of secondary structure, ranking of best silencing sequences, etc.

Another aspect of the invention is a computer program as described herein, further comprising the ability to provide results as a written file, as an email message, posted onto the Internet or provided by any other means known in the art for making results available. Where the method is implemented into a workstation environment or on a web server, for example, the user may set up a method of the present invention and log out of a session or close a browser window, and the method will run in the background. The results may be made available upon completion. The option to enter such offline mode may be provided at one or more stages in a session. For example, steps b) and c) may be executed offline. Alternatively, where allowable mismatches are not checked (*e.g.* G/T rule), the steps b) and c) may be executed online, and the best RODI resubmitted to the method in an offline mode where allowable mismatches are checked and the results emailed to the user.

Another aspect of the invention is a computer program as described above further comprising one or more databases of gene sequences of the SOI.

Another embodiment of the present invention, is a computer program as described herein, further comprising the ability of displaying said user interface on a web browser of a remote computer connected to the Internet.

Another embodiment of the present invention is a computer readable medium on which is stored a computer program capable of performing a method of the present invention. Said computer readable medium may be any, such as, for example, an optical disc (CD-ROM, DVD-ROM), a solid-state ROM/RAM, a floppy disk.

Kit

Another embodiment of the present invention is a kit comprising at least one computer readable medium on which is stored a computer program capable of performing a method of the present invention.

Another aspect of the invention is a kit further comprising one or more vectors suitable for use in modulating a GOI and/or FGOI. Said vectors may be in an open or closed form.

According to one aspect of the invention, suitable vectors are those which are capable of producing dsRNA in the SOI.

According to another aspect of the invention, suitable vectors allow the construction of a clone comprising tandemly-arranged (+) and (-) strands as described above. According to one aspect of the invention, vectors are designed to allow so-called recombinational cloning in one step. This may be achieved, for example, by amplifying the sequence of interest with PCR primers extended to include recombination sites, and mixing the so amplified DNA with the appropriate vectors and enzyme mix.

Suitable vectors and technologies are known in the art, and are described in WO 02/059294 and WO 99/53050 which are incorporated herein by reference. Examples of vectors include pDON, pHellsgate, pHellsgate 8, pHellsgate 11, and pHellsgate 12 as described in WO 02/059294.

According to another aspect of the invention, suitable vectors are any comprising one or more promoters for transcription of the DNA of interest, and one or more 3' end regulatory regions such as, for example, transcription termination and polyadenylation signals.

Clones so formed may be introduced into a SOI using any method. For example, if the sequence of interest is cloned into a T-DNA vector, the *Agrobacterium* system may be used to introduce the gene into the plant cell.

According to another aspect of the invention, a suitable vector comprises a bacteriophage promoter. In this case, dsRNA is generated by *in vitro* transcription and said RNA is introduced into the SOI.

Another aspect of the invention is a kit further comprising one or more reagents for cloning tandemly-arranged (+) and (-) strands of modulating DNA identified according to the invention into a vector. Examples of reagents include any of proteins for recombination (*e.g.* INT, XIS, IHF, clonase™), buffers.

Another aspect of the present invention is a kit further comprising a computing device.

Another embodiment of the present invention is data obtained from a method of the invention.

Another embodiment of the present invention is a vector comprising a sequence obtained using data from a method of the invention.

Having understood the currently described methods, a person skilled in the art will immediately realise that the methods of the invention can equally be applied to determine the optimal gene-silencing RNAi or DNA yielding such RNAi.

According to one variation, a chimeric gene may be introduced in a first (principal) organism with the ultimate goal to effect gene silencing of a target gene in another (secondary) organism which is capable of receiving the chimeric gene, RNAi, or part thereof from principal organism. Applications of such a method include, for example, expressing RNAi or dsRNA encoding a chimeric gene in a plant, whereby the RNAi produced in the plant is targeted to silence a gene in a pest feeding on such a plant.

In one aspect of the invention the GOI is compared, according to the described methods, against the genome or transcriptome of the secondary organism from which the GOI is derived. The RODI obtained therefrom, may then be further applied to a method of the invention to determine whether said RODI silences any genes in the principal organism *i.e.* the organism from which the GOI is not naturally comprised and in which the silencing RNA may initially be expressed. The resulting RODI are specific for the GOI in the secondary organism, and are divergent from other genes in both the principal and secondary organisms. In this instance, there are two SOIs – one naturally devoid of the GOI, and one comprising the GOI.

According to another aspect of the invention, the GOI is compared, according to the described methods, against the combined genomes or transcriptomes of the principal and secondary organisms. The resulting RODI are specific for the GOI in the secondary organism, and are divergent fróm other genes in both the principal and secondary organisms. In this instance, the SOI comprises both the primary and secondary organisms.

The RODI may be further processed to rank them according to predicted suitability for silencing the expression of the GOI, while not silencing the expression of any other gene in the SOI.

To this end, the RODI may be analyzed to determine those RODI comprising the smallest number of windows of at least 12 nucleotides as previously described, and preferably of windows of 19 nucleotides, wherein the number of mismatches is low (*i.e.* wherein the number of mismatches is 1, 2, 3, 4 or 5).

Thus, as an example, a RODI wherein, for example, the highest level of sequence identity with the SOI is a window of 19 nucleotides having at least one mismatch, will be ranked lower in suitability than a RODI wherein the highest level of sequence identity is a similar sized window having more than one mismatch.

DETAILED DESCRIPTION OF THE FIGURES

Figure 1 is a flow chart of part of a method according to the invention. According to the embodiment shown in Figure 1, a database of the species of interest, SOI, **1**, is accessed and the sequence of first gene therein, GENE *n*, is retrieved. The coding strand of the gene of interest, GOI(+), **2**, is also provided. The sequences of GENE *n* and GOI(+) are compared **3**. If GENE *n* and GOI(+) are statistically similar **4**, the name or the sequence of GENE n is stored in list C^p, **5**, and *m* in incremented by 1. The next gene in the database is accessed by incrementing gene counter *n* by 1, **6**. The methods of boxes **3**, **4**, **5** and **6** are repeated for the new gene. The cycle is repeated until all *n* has counted all the genes of the SOI in the database **7**. The list C^m is parsed, and genes of the SOI having an identity of less than 15 nucleotides in common with the GOI are discarded **9**. The remaining gene list, C^m, **10** is carried forward to the next stage of the method as shown in Figure 2.

Figure 2 is a flow chart of part of a method according to the invention. According to the embodiment shown in Figure 2, the first nineteen bases of the coding strand of the GOI, *i.e.* when position *p* =1, are stored in variable $GOI(+)19^p$ **21**. The first nineteen bases of the antisense strand of the GOI, *i.e.* when position *p* =1, are also stored in variable $GOI(-)19^p$ **22**. Each 19mer of $GOI(+)19^p$ or $GOI(-)19^p$ is compared against the list of genes obtained in box **10** of Figure 1 as follows. The sequence of first gene obtained from Figure **1**, *i.e.* C^m when *m* =0 , is checked for identical occurrences of sequence $GOI(+)19^p$, with one mismatch allowed *i.e.* stringency = 18 nucleotides in box **23**. Similarly, the sequence of C^m is checked for identical occurrences of sequence *GOI(-)19^p^*, with one mismatch allowed *i.e.* stringency = 18 nucleotides in box **24**. Where there is identity within the allowed stringency (**25**, **26**), said 19 nucleotide identity region is indicated on the sequence of the GOI, and recorded in this embodiment as *GOI(+)mrk*, **211** or *GOI(-)mrk*, **212**. The next gene of the list obtained in Figure 1 is retrieved in this embodiment by incrementing *m* by 1, **27**. The steps of boxes **23**, **25**, **211** and **27** for the coding (+) strand, and the steps of boxes **24**, **26**, **212** and **27** are repeated for the antisense (-) strand, until the list of genes obtained in Figure 1 is exhausted, **28**. When the list of genes from Figure 1 is exhausted, the method examines the next 19 nucleotides in the GOI of the coding and antisense strands by, in this embodiment, increasing counter *p* by 1, and resetting the counter m, **29**. As the 19 nucleotide window is moved across the GOI of interest, an indication of identity within the required stringency, with the genes identified in Figure 1 is marked in *GOI(+)mrk* **211** or *GOI(-)mrk* **212**. Thus *GOI(+)mrk* or *GOI(-)mrk* accumulate indications of identity until the 19 nucleotide window has passed across the full sequence of the GOI, whereupon the loop ends **213**. The output from the method is carried forward to a next stage of the method, shown in Figure 3.

Figure 3 is a flow chart of part of a method according to the invention. According to the embodiment shown in Figure 3, the regions of identity with the required stringency **31**, **33**, identified in Figure 2 are subtracted from the sequence of the GOI **32**, to provide an indication of the regions of divergent identity (RODI) in the GOI **34**. The output therefrom **35**, may be used according to the method to provide various lists and rankings. Examples of such lists include:

- sequences of regions of divergent identity which are equal to or longer than 300 nucleotides in continuous sequence, **36**,
- sequences of regions of divergent identity which are shorter than 300 nucleotides in continuous sequence, **37**,
- regions of divergent identity which are shorter than 300 nucleotides in continuous sequence, combined to provide a new concatenated sequenced greater than 300 nucleotides in length **38**,
- sequences of regions of divergent identity wherein regions of secondary folding and/or degree of secondary folding is indicated, **39**,
- proposed sequences of RNA duplexes suitable for modulating the GOI, **310**.

Figure 4 is a flow chart of part of another embodiment of a method according to the invention. According to the embodiment shown in Figure 4, the coding strand of the gene of interest, GOI (+), **41**, is used to make BLAST computation, **43**, against a database of the species of interest, SOI, **42**.

The BLAST output is filtered, **44**, to retrieve all the genes that share at least 15 nucleotides identity with GOI (+). These genes are stored in candidate list C^m, **45**, that is carried forward to the next stage of a method, shown in Figure 5.

Figure 5 is a flow chart of part of a method according to the invention. According to the embodiment shown in Figure 5, the sequence of the first candidate gene, Candidate n, **53**, identified by the scheme in Figure 4 is retrieved from database of the SOI, **51**, with seqret, a program from EMBOSS package, **52**. The sequence of Candidate n is checked by the COMPARE program, **55 & 512**, for identical occurrences of 19 nucleotides and one mismatch allowed (*i.e.* stringency = 18 nucleotides) with sequence GOI (+), **54**, and the reverse sequence GOI (-), **511**, obtained with revseq, a program from EMBOSS package, **510**. If there is identity within the allowed stringency, **56** & **513**, said 19 nucleotides identity region, it is indicated on the sequence of the GOI, recorded in this embodiment as GOI (+) mrk, **57**, and the gene name of the Candidate n sequence is stored in cross silencing gene list CSGp, **57**. For the GOI (-), the identity region position is converted to the strand of GOI (+), **514**, in order to record it in GOI (+) mrk. The next sequence is obtained by incrementing n by 1, **58**, until the end of candidate list is exhausted, **59**, whereupon the loop ends, **515**. The output, GOI (+) and CSGp are carried forward to the next stage of a method, shown in Figure 6.

Figure 6 is a flow chart of part of a method according to the invention. According to the embodiment shown in Figure 6, the sequence of the first candidate gene, Candidate n, **63**, is retrieved from SOI, **61**, with seqret **62**, a program from EMBOSS package. A domain (GOIDd) **65** and **615** of the gene of interest of 21 nucleotides in length is extracted from position d (initially 1) the start of the GOI (+) **64** and (-) **614** sequence. A check is made **66** and **616** to test whether the GOIDd is within the GOI and not outside. If it is not outside, the sequence of Candidate n **63** retrieved by seqret **62** is checked by FUZZNUC **67**, **617**, a program from the EMBOSS package, for the presence of GOIDd **65**, **615** with one mismatch allowed, and patterns defining T as T or C, this solution allowing GT pairing in the DNA level to mimic the GU pairing at the level of RNA. If GOIDd is found within the allowed parameters, **69**, **619**, it is indicated on the sequence of the GOI, recorded in this embodiment as GOI (+) mrk, **610**, and the gene name of the Candidate n sequence is stored in cross silencing gene list CSG^{p} **610**. For the GOI (-), the identity region position is converted to the strand of GOI (+), **620**, to record it in the GOI (+) mrk. A new GOIDd is obtained by shifting the domain to the right of one nucleotide, **68**, **618**. Once the 21nt domain has reached the end of the GOI, (**66**, **616**, YES) next sequence of Candidate gene is obtained by incrementing n by 1, **611**, until the end of candidate list is exhausted, **612**, whereupon the loop ends, **621**.

Figure 7 is a flow chart of part of a method according to the invention. According to the embodiment shown in Figure 7, the regions of identity with the required stringency, **71**, identified in Figures 5 and 6 are subtracted from the sequence of the GOI, **72**, to provide an indication of the regions of divergent identity (RODI) in the GOI, **73**.

The output therefrom **74**, may be used according to the method to provide various lists and rankings. Examples of such lists include:

- list of cross silencing genes obtained in figures 5 or 6, **75**
- regions of identity N-masked on the GOI (+), **76**
- regions of identity in lower case on the GOI (+), **77**
- list of all RODI present on the GOI, **78**
- list of best RNA duplexes (maximum 3 sequences) for modulating GOI, **79**.

Figure 8 is a flow chart of part of a method according to the invention. According to the embodiment shown in figure 8, all RODI obtained in figure 7 are ranked by decreasing length, **81** in a list with n members. The indices n & v are set to **1**, **82**, and by this fact the longest sequence of RODI is processed first, **83**. If the length of RODI^{n}, **83**, is longer than 300 nucleotides, **84**, RODI^{n} is added to the list of best RNA duplexes for modulating GOI, **85**. The next RODI is checked by incrementing n & v by 1, **88**, until the 3 best RNA duplexes are obtained, **86**, or all RODI are checked, **87**.

If the length of RODI^{n} is smaller than 300 nucleotides, **84**, we enter in a loop to construct a RODI ($\sum x=v\rightarrow n$ RODIx) of the required length by concatenation of several smaller RODI. Until the length of the constructed RODI is longer than 300 nucleotides, **811**, a new smaller RODI, **89**, is added to the construction, **810**. When $\sum x=v\rightarrow n$ RODIx reaches the correct length ($\geq$300 nt), **811**, the small RODI are reorganized according to their original position on the GOI, **812**, and stored in the list of best RNA duplexes for modulating GOI, **813**, and the loop is exited, **814**.

Figure 9 depicts a flow chart of a method the invention. The sequence of GOI is inputted, along with the options of program **91**. Form the input sequence **92**, a BLAST search is performed **93**, taking as input the database of the SOI **94**. The output of BLAST is parsed **95**, and genes which contain an identity longer than 15 nucleotides are filtered **96 & 97**, and the names are stored in a candidates list **99**. Those which do not are rejected **98**. A detailed example of performing steps **91** to **99** is provided in Figure 4.

The sequence of each gene of the candidate list C^m is extracted from the SOI database **94** by seqret **910** and compared with the coding strand of the input sequence, GOI(+), **92** using COMPARE algorithm from GCG **911**. At the same time, the antisense strand, GOI(-) **914** of the input sequence, **92** is created by revseq **913**, and compared with the candidate gene sequence also using COMPARE algorithm from GCG **915**. The output **912** from both COMPARE routines are parsed **916** and the localisation procedure **917** marks the regions of identity on the GOI. A detailed example of performing steps **910** to **917** is provided in Figure 5.

The procedure **918** extracts the RODI from the GOI and the classification procedure **919** reorganizes these regions to create the output results files **920**. Detailed examples of steps **918** to **920** are provided in Figures 3 and 6.

Figure 10 depicts a flow chart of a method the invention. The sequence of GOI is inputted, along with the options of program **101**. Form the input sequence **102**, a BLAST search is performed **103**, taking as input the database of the SOI **104**. The output of BLAST is parsed **105**, and genes which contain an identity longer than 15 nucleotides are filtered **106**, **107**, and the names are stored in a candidates list **109**, and are not rejected **108**. Steps **101** to **109** are essentially equivalent to the steps detailed in Figure 1.

The sequence of each gene of the candidates list **109** is extracted from the SOI database **104** by seqret **1010** and compared with the GOI(+) **102** using COMPFUZ algorithm **1011** which is equivalent to FUZNUC combined with a position incrementer (see Figure 6 dotted rectangle, **65** to **68**). At the same time, the antisense strand **1014** of the GOI **102** is created by revseq **1013**, and compared with the candidate gene sequence also using COMPFUZ algorithm **1015**. The output **1012** from both COMPFUZ algorithms is parsed **1016** and the localisation procedure **1017** marks the regions of identity on the GOI. Steps **1010** to **1017** are essentially equivalent to the step detailed in figure 3.

The last steps (1018 to 1020) are essentially equivalent to the steps described in Figures 3 and 6. The procedure **1018** extracts the RODI from the GOI and the classification procedure **1019** reorganises these regions to create the output results files **1020** which can be, for example, written to a networked computer, emailed, or published on the Internet.

Figure 11 depicts a flow chart of a method of the invention wherein each member of family of gene of interest (FGOI) **111**, is aligned and a single homologous sequence is derived therefrom **112**. The sections of the members which exhibit homology with each other are indicated as hatched boxes. The sections of the single homologous sequence which show little/no homology are excised, and the homologous sections are concatenated to form a single GOI, **113**, for input into a method of the present invention. Alternatively, each of the sections of the single homologous sequence which show homology are separated used as a single GOI, **114**, for input into a method of the present invention.

Figure 12 depicts a flow chart of a method of the invention wherein the choice of silencing a single gene of a family of genes is indicated by the user, along with names or sequences of said gene(s) **121**. Based on the choice **122**, where a family of genes is to be silenced, the sequences or names of the genes are obtained **123**, and using the Clustal W routine **124**, a single consensus sequence is generated **125** which is used as an input sequence **126** for determining the ROI **128**. Where the silencing of a single gene is chosen **122**, the sequence of said gene is obtained **127** and used as an input for determining the RODI **128**. The steps of determining the RODI **128**, generating a list of RODI **129** and output results **1210** is described extensively elsewhere herein.

Figure 13A: depicts another flow chart of a method of the invention, in which a family of genes is to be silenced. The sequences of genes of family (FGOI) are inputted, along with the options of program **1311**. From the input sequences **1312**, a COMPARE is performed for each genes and two by two, **1313**. The output of each COMPARE is parsed, **1314**.

For each output parsed, **1315**, the number of common words between two genes is calculated (CW score), **1317**. If no hit found between two genes, the "WO match" score (without hit) is increased by 1, **1316**. The results are stored in the list FGOI^R, **1318**.

For each gene of the family, the number of common words is calculated and stored in a table, **1319**.

The matrix can be showed on a web page and the best representative gene (BRG) is chosen by the greatest CW score with the smaller "WO match" score, **1320**. At this step, the user can choose another gene. Only one gene can be submitted as the representative of the family. The BRG is inputted in the classical program described before (*e.g.* Figures 2, 3, 5, 6).

Figure 13B: depicts another flow chart of a method the invention for Family part. The sequences of genes of family (FGOI) are inputted, along with the options of program **1311**. From the input sequences **1312**, a COMPFUZ **1321** routine is performed *i.e.* FUZZNUC combined with a routine to increment the domain window by 1 nt. COMPFUZ is essentially similar to the steps described in Figure 6 (dotted rectangle, **65** to **68**). COMPFUZ **1321** is performed for all possible pairs of genes, two by two,. The output of each COMPFUZ calculation is parsed, **1314**.

For each output parsed, **1315**, the number of common domains between two genes is calculated (CD score), **1317**. If no hit found between two genes, the WO score (without hit) is incremented by 1, **1316**. The results are stored in the list FGOI^R, **1318**.

For each gene of the family, the number of common domains is calculated and stored in a table, **1319**. The matrix may be shown in a web page and the best representative gene (BRG) is chosen by the greatest CD score having the smaller WO score, **1220**. At this step, the user can choose another gene. Only one gene can be submitted as the representative of the family. The BRG is inputted in the classical program described before (*e.g.* Figures 2, 3, 5, 6).

Figure 14 is a flow chart of another option of the invention. According to the embodiment shown in figure 14, each coding strand of the genes of family, FGOI (+) **141** is compared two by two. A first gene of interest GOI^n (+) **143** of the family is obtained using index n, 142, and a second gene of interest GOI^m (+), **1410**, is obtained using the index m, m being a higher index than n, **148**.

A domain of 19 nucleotides (GOIDd) **144** is extract from the GOI^n (+) sequence **143**. If GOIDd is within the GOIn sequence, and not outside **145**, the sequence of GOIm **1410** is checked by FUZZNUC **146** for the presence of GOID^d with one mismatch allowed and patterns defining T as T or C. If GOID^d is found within the allowed parameters **1411** the number of common domains (CD^{nm}) between these two genes is incremented by 1, **1412**. If no hit found between two genes, the WO^{nm} score (without hit) is incremented by 1, **1413**. The results are stored in a matrix, **1414**. A new $GOID^{d}$ is obtained by shifting the domain to the right of one nucleotide, **147**. The next sequence of GOIm is obtained by incrementing m by 1, **1415**, until the end of genes family list is exhausted, **149** (YES). The next sequence of GOIn is obtained by increasing n by one, **1416**, until the end of candidate list is exhausted, **1417** (YES), whereupon the loop ends, **1418**. The best representative gene (BRG), **1419**, is chosen by the greatest CD score with the smaller WO score. At this step, the user can choose another gene. The best representative gene (BRG) is the GOI used in the method described above.

Figure 15 depicts an example of a user interface according to a computer program of the invention capable of display a web browser on a remote computer. At the start of the program the user may indicate the following parameters:

- the GOI by sequence, accession number of database name **151**,
- which database of the biological system or species of interest (SOI) is used **152**,
- whether to silence a single gene or a family of genes **153**,
- a job name **154**,
- the minimum **155** and maximum **156** lengths of fragment,
- the wordsize **157** and stringency **15** for the COMPARE algorithm,
- the output format of sequences: number of nucleotides per line **159** and the insertion of a space between a given number of nucleotides **1510**.

Figure 16 depicts an example of a user interface according to a computer program of the invention capable of display a web browser on a remote computer. While the program is performing the BLAST and COMPARE routines, it may indicate its status and provide a tabulated summary of the input parameters **161**.

Figure 17-1 depicts an example of a user interface according to a computer program of the invention capable of display a web browser on a remote computer. Once the program has finished running, it may first present a summary of the input parameters **171**, and the sequence of the GOI, wherein regions of identity with genes of the SOI are indicated by an N, **172**.

Figure 17-2 at the bottom of the same summary screen, the program may give several options for viewing the results of the search **173**, such as "lower case sequence" (identity regions shown in lower case), "N-marked sequence" (identity regions shown as an N), "cross-silencing genes" (lists genes which showed identity with the non-divergent regions of the GOI), "BEST sequence" (display a sequence meeting the parameters of the user, *e.g.* between 300 to 800 nucleotides in length), "Good sequence" (displays all the regions of the GOI which have divergent identities from the SOI).

Figure 18 depicts an example of a user interface according to a computer program of the invention capable of displaying a web browser on a remote computer. Shown here is an example of the display option "good sequence" which may list all the sequences of the SOI, regardless of length, which have divergent identities from the GOI.

Figure 19 depicts an example of a user interface according to a computer program of the invention capable of display a web browser on a remote computer. Shown here is an example of the display option "BEST sequence" which may list display a sequences meeting the parameters of the user, *e.g.* between 300 to 800 nucleotides in length. In this case, no single region met this requirement, and the program provided a sequence of the required length **192** by concatenating several smaller sequences **191**.

Figure 20 depicts an example of a user interface according to a computer program of the invention capable of display a web browser on a remote computer. Shown here is an example of the display option "cross-silencing genes" which may list all the genes of the SOI which showed identity with the non-divergent regions of the GOI.

Figure 21 depicts an example of a user interface according to a computer program of the invention capable of display a web browser on a .remote computer. Shown here is an example of the display option "lower case sequence" in which the identity regions of the GOI are shown in lower case, and the divergent regions are shown in upper case.

Figure 22 depicts an example of a user interface according to a computer program of the invention, displaying the result of a search for RODIs within a family of genes. Displayed is the query sequence **221**, represented as a graphical bar **222**, which indicates shaded regions according to the key **228**. Below are members of the FGOI **223** on which is indicated regions of homology within the family **225**, and regions of homology within the FGOI **226** considered too short. Below there are genes of the SOI **224**, and an indication given of sequences which must not be silenced **227**.

Figure 23 depicts an example of a user interface according to a computer program of the invention, displaying the best sequence **231** for silencing the target gene resulting from a method of the present invention, without using the G/T rule. It provides the option **232**, for submitting the best sequence, rather than the whole gene to the present method which implements the G/T rule.

Figure 24 depicts an example of a user interface according to a computer program of the invention capable of display a web browser on a remote computer. Shown here is an output form indicating a matrix of genes to be silenced from a family. The sequence numbers are indicated along the edges **241, 242**, scores for each pair, the total score (CD) for a gene **243** and the WO match **244** are also shown. The best representative match is selected as the gene with the largest score and the smallest WO match **245**.

CLAIMS

1. A method for identifying nucleic acid capable of modulating a specific gene of interest, GOI in a biological system of interest, SOI, comprising:

(a) obtaining a sequence of the GOI,

(b) identifying genes in the SOI that share a nucleotide identity with the GOI,

(c) determining one or more regions of the GOI that have a divergent identity with the genes identified in step (b), and

(d) identifying, from the regions determined in step (c), nucleic acid capable of modulating a specific GOI in a SOI.

2. A method according to claim 1 wherein a gene of said SOI of step (b) is identified by determining whether any window of at least 12 nucleotides of one strand of the GOI, exhibits identity to said gene of the SOI.

3. A method according to claim 1 or 2 wherein step (c) further comprises the following steps:

(c1) determining which windows of at least 12 nucleotides in length of both strands of the GOI exhibit identity with any of the genes of the identified in step (b), said identity permitting at least one mismatch, and

(c2) providing regions of the GOI devoid of said nucleotide windows identified in step (c1).

4. A method according to any of claims 1 to 3 wherein the identity in step (b) is between 17 and 25 nucleotides.

5. A method according to claims 3 and 4 wherein the window in step (c1) is between 17 and 25 nucleotides, and the number of mismatches is 1.

6. A method according to any of claims 1 to 5, wherein the step (b) comprises the use of a statistical evaluation of homology.

7. A method according to any of claims 1 to 6, wherein the step (b) comprises the use of the "BLAST" algorithm or variant thereof.

8. A method according to any of claims 1 to 7, wherein the step (c) comprises the use of the "COMPARE" algorithm from the GCG package, or a variant thereof.

9. A method according to any of claims 3 to 8, wherein step (c1) treats sequences which are non-identical, but can hybridise to the same oligonucleotide *via* G:U or G:T mismatching as exhibiting identity.

10. A method to claim 9, wherein said non-identical sequences are identified using the "NUCFUZZ" algorithm from the EMBOSS package, or a variant thereof.

11. A method according to any of claims 1 to 8 wherein the best region identified in step (d) is submitted to the method according to claims 9 or 10, so the G/T rule is implemented after the best sequence has been found.

12. A method according to any of claims 1 to 11 further comprising a step of concatenating two or more regions of the GOI determined in step (c) when the maximum length of a region determined in step (c) is less than a minimum number of nucleotides determined by the user.

13. A method according to any of claims 1 to 12 further comprising a step of determining the sequence of at least one duplex RNA, or stem-loop wherein one strand of said RNA duplex or stem corresponds to at least part of a divergent region determined in step (c).

14. A method according to claim 13 comprising the step of determining potential secondary structure-forming regions of said duplex RNA.

15. A method according to any of claims 1 to 14 further comprising a step of determining at least one DNA sequence suitable for cloning into a vector, wherein said DNA sequence corresponds to at least part of a divergent region determined in step (c).

16. A method according to claim 15, wherein said vector is capable of producing double stranded RNA in the SOI.

17. A method according to any of claims 13 or 14 wherein said RNA duplex comprises at one substitution where U is C, C is U, G is A, or A is G.

18. A method according to any of claims 1 to 17 further comprising a step of determining the sequences of at least one pair of PCR primers, suitable for amplifying the DNA sequence(s) of claim 11.

19. A method according to any of claims 3 to 18, wherein
- steps c1) to d) are repeated, and
- the number of mismatched permitted in step c1) is increased by one after each repeat,
so producing a list regions of the GOI that have a divergent identity with the genes identified in step (b), corresponding to an increasing number of permitted mismatched.

20. A method for identifying nucleic acid capable of modulating a family of genes in a biological system of interest, SOI, comprising the steps of:

(A) obtaining the sequences of the genes in the family of genes of interest, FGOI,

(B) calculating a single homologous sequence from the sequences of step (A),

(C) selecting each region of homology calculated in step (B) as a GOI,

(D) identifying nucleic acid capable of modulating the FGOI by using each GOI of step (C) in a method according to any of claims 1 to 19.

21. A method according to claim 20, wherein the regions of homology in step (C) are concatenated to form a single GOI.

22. A method according to claims 20 or 21, wherein a single homologous sequence of step (B) is calculated using Clustal W.

23. A method according to claims 20 or 21, wherein a single homologous sequence of step (B) is a best representative sequence, BRG, calculated by comparing genes of the FGOI a pair at a time.

24. A method according to claim 23, wherein the BRG is calculated by:
(I) selecting a sequence of 18 to 20 nt from the start of a first gene of the FGOI,

(II) comparing said sequence across each of the other genes of the FGOI for an identity match, or no identity match,

(III) repeating step (I) using the next gene of the FGOI, until all the genes of FGOI have been exhausted, and

(IV) selecting the gene with the highest number of identity matches and the lowest number of no identity matches, which is the BRG .

25. A computer program stored on a computer readable medium, capable of performing a method according to any of claims 1 to 24.

26. A computer program according to claim 25 comprising an ability to display a user interface on a computer, said interface allowing a user to provide one or more parameters to a method of the invention.

27. A computer program according to claims 25 or 26 further comprising an ability to display a user interface on a computer, said interface allowing a user to receive an indication of the sequences provided by a method of the invention.

28. A computer program according to any of claims 25 to 27 further comprising an ability to make available the sequences provided by a method of the invention by email, by Internet publishing, or by storing on a networked computer.

29. A computer program according to any of claims 25 to 28 further comprising one or more databases of gene sequences of the SOI.

30. A computer program according to any of claims 25 to 29, further comprising an ability to display said user interface on a web browser of a remote computer connected to the Internet.

31. A computer readable storage device on which a computer program according to any of claims 25 to 30 is stored.

32. A kit comprising at least one computer readable storage device according to claim 31 and one or more vectors suitable for use in gene modulation.

33. A kit according to claim 32 wherein said vectors are any capable of producing double stranded RNA in the SOI.

34. A kit according to claim 33 wherein said vectors are one or more of pDON, pHellsgate, pHellsgate 8, pHellsgate 11, and pHellsgate 12.

35. Unknown nucleic acid identified or determined according to a method of any of claims 1 to 24.

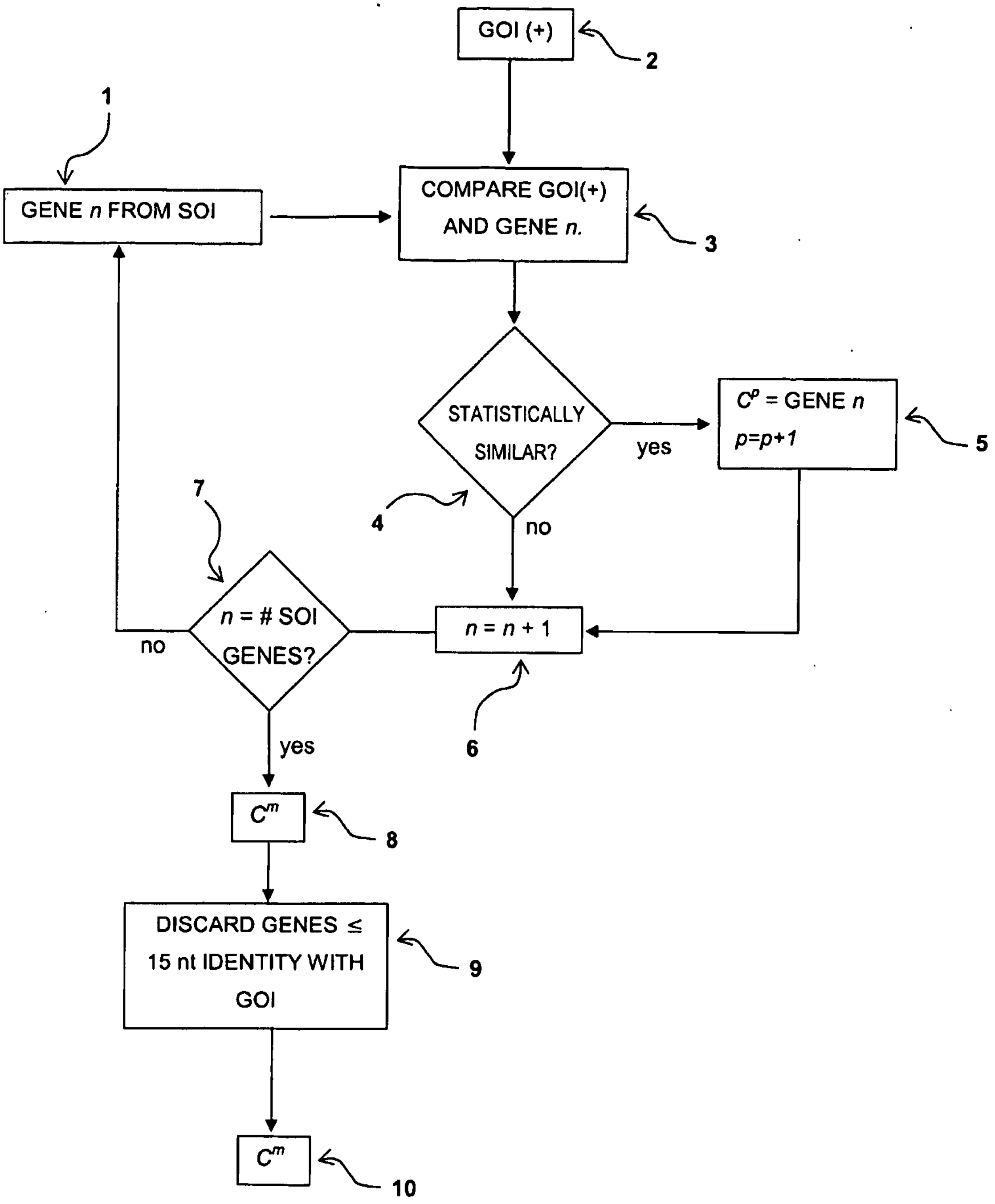

GOI (+)
2
1
GENE *n* FROM SOI
COMPARE GOI(+) AND GENE *n*.
3
STATISTICALLY SIMILAR?
yes
C^p = GENE *n*
p=*p*+1
5
4
no
7
n = # SOI GENES?
no
n = *n* + 1
6
yes
C^m
8
DISCARD GENES ≤ 15 nt IDENTITY WITH GOI
9
C^m
10

FIGURE 1

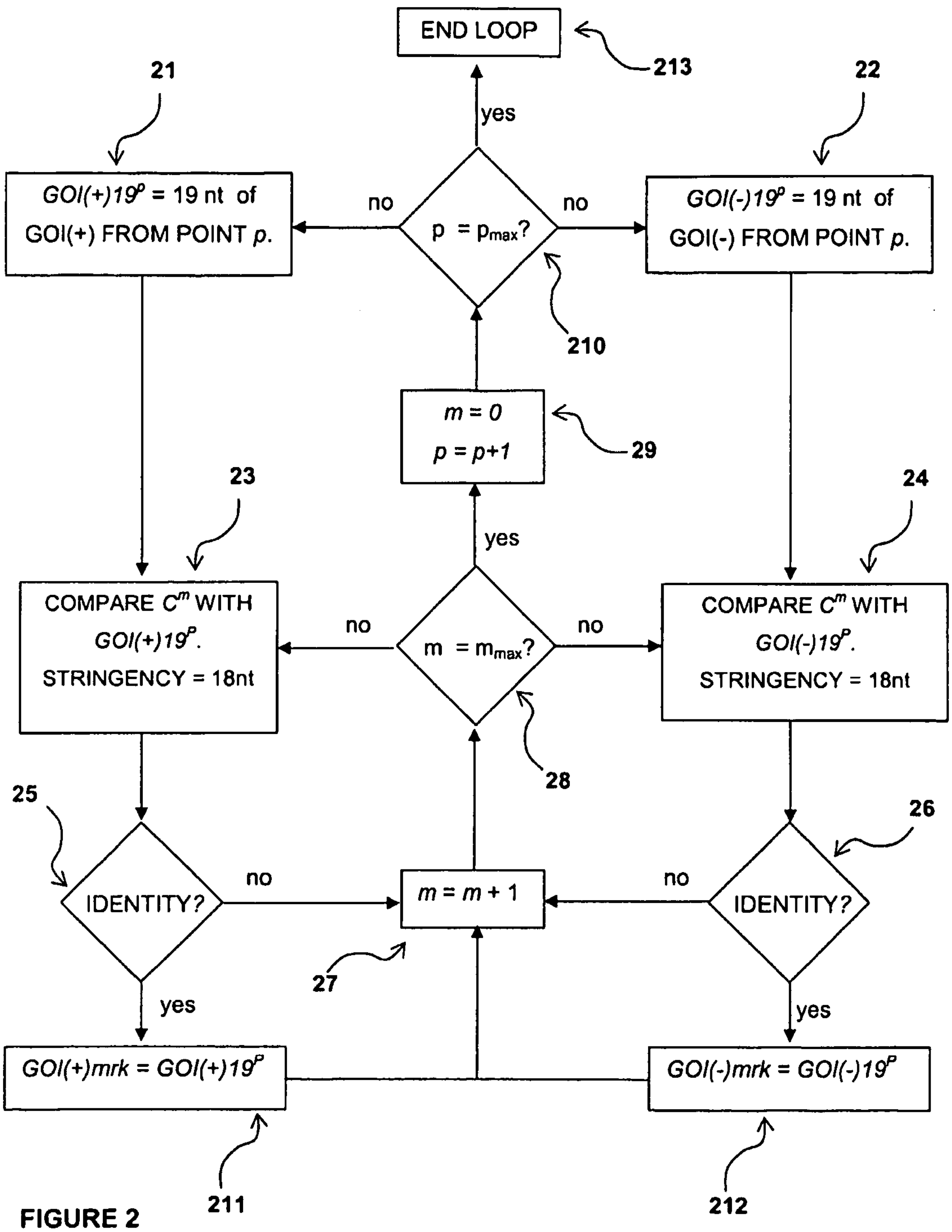

END LOOP
213
21
22
yes
no
no
$GOI(+)19^p$ = 19 nt of
GOI(+) FROM POINT p.
p = p_{max}?
$GOI(-)19^p$ = 19 nt of
GOI(-) FROM POINT p.
210
$m = 0$
$p = p+1$
29
23
24
yes
COMPARE C^m WITH
$GOI(+)19^P$.
STRINGENCY = 18nt
no
m = m_{max}?
no
COMPARE C^m WITH
$GOI(-)19^P$.
STRINGENCY = 18nt
28
25
26
IDENTITY?
no
$m = m + 1$
no
IDENTITY?
27
yes
yes
$GOI(+)mrk = GOI(+)19^P$
$GOI(-)mrk = GOI(-)19^P$
211
212

FIGURE 2

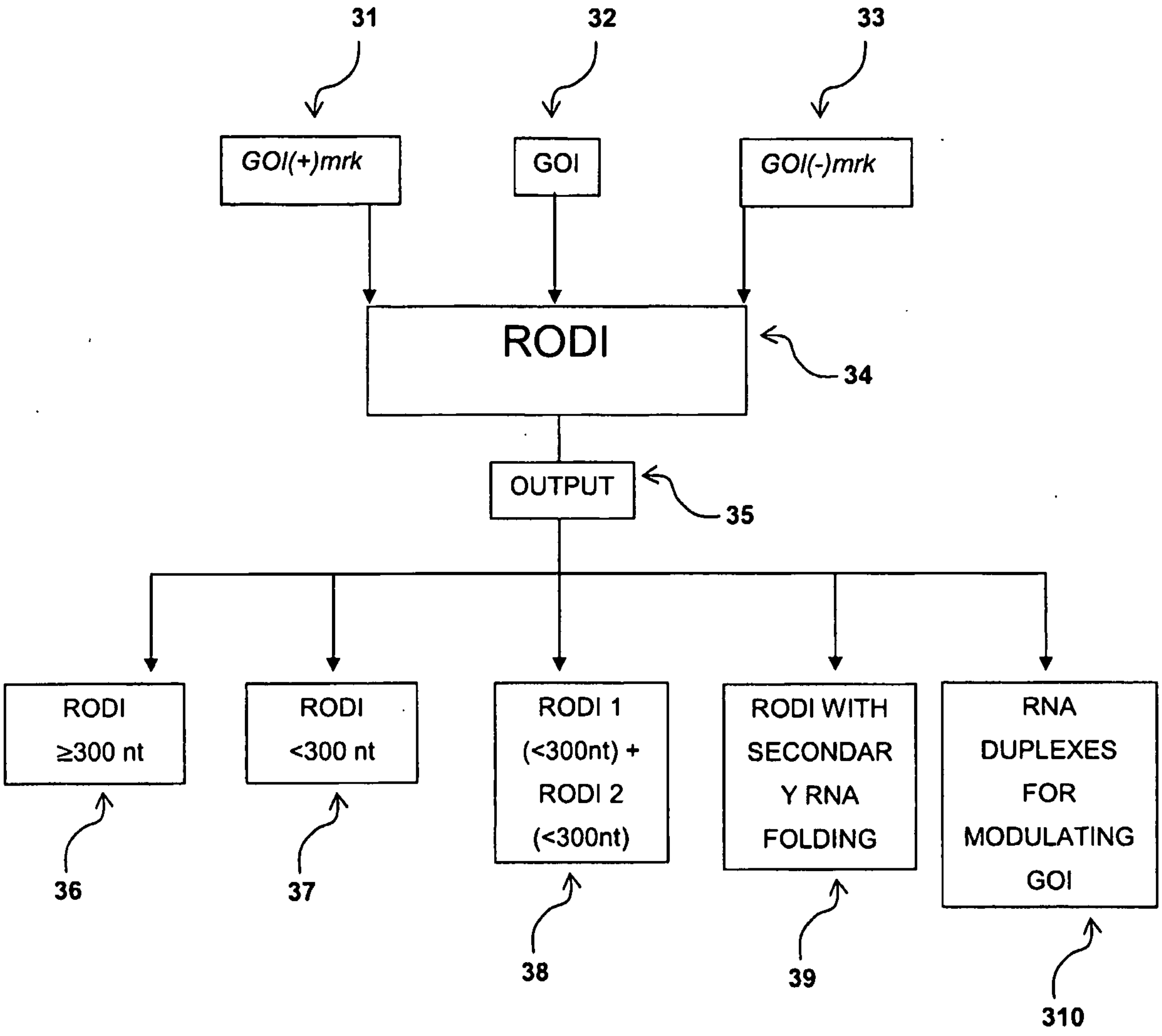

31
32
33
GOI(+)mrk
GOI
GOI(-)mrk
RODI
34
OUTPUT
35
RODI ≥300 nt
RODI <300 nt
RODI 1 (<300nt) + RODI 2 (<300nt)
RODI WITH SECONDARY RNA FOLDING
RNA DUPLEXES FOR MODULATING GOI
36
37
38
39
310

FIGURE 3

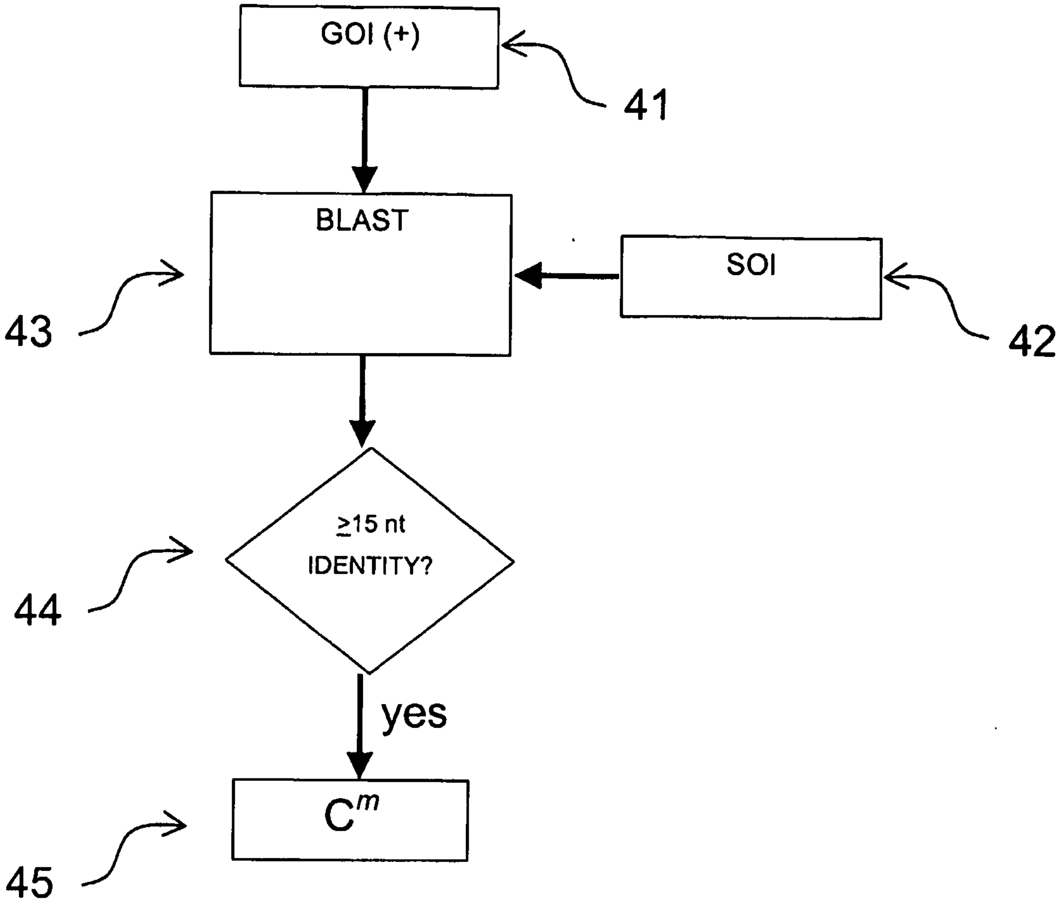


FIGURE 4

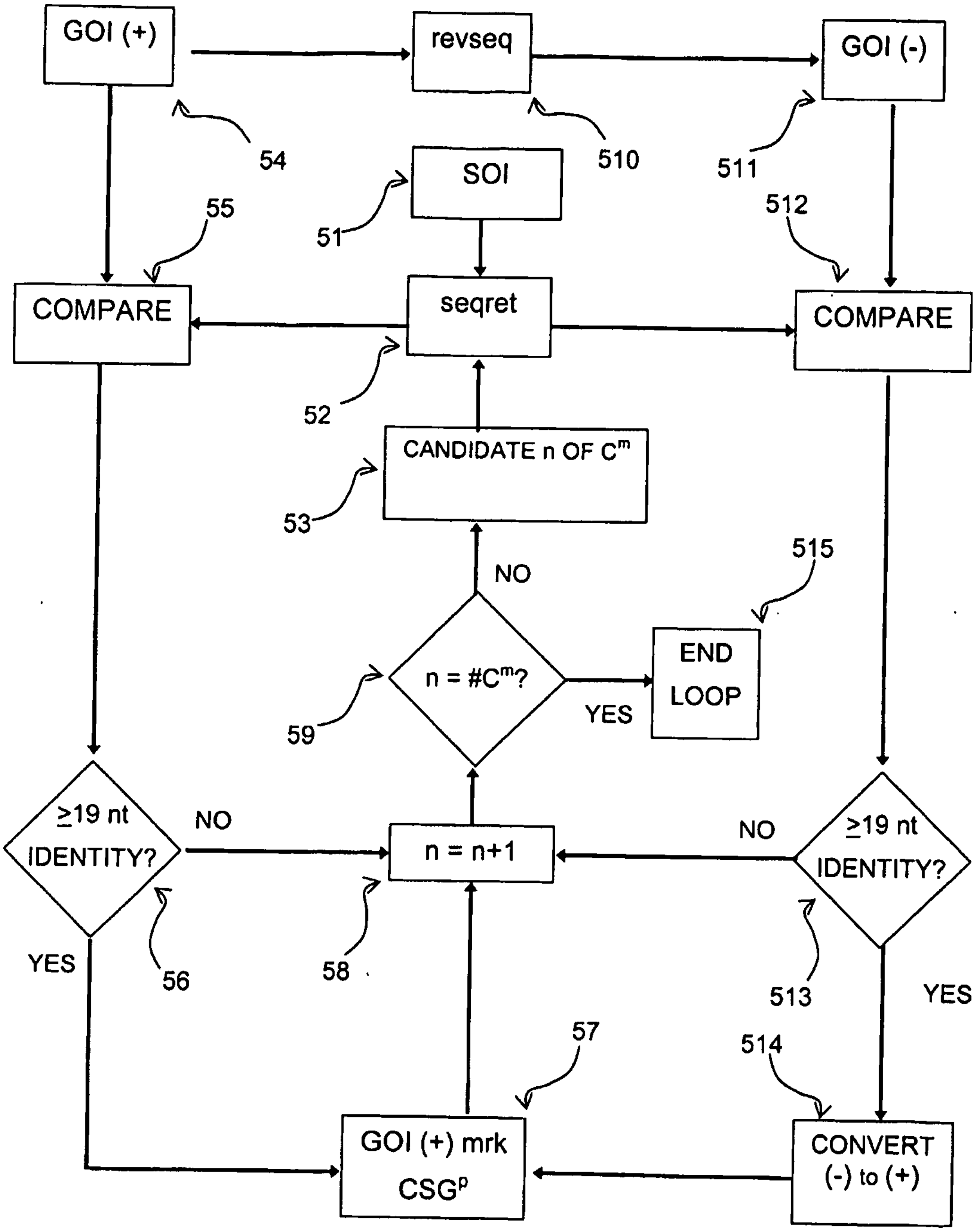

GOI (+)
revseq
GOI (-)
54
510
511
SOI
51
55
512
COMPARE
seqret
COMPARE
52
CANDIDATE n OF C^m
53
515
NO
END LOOP
$n = \#C^m$?
YES
59
≥19 nt IDENTITY?
NO
n = n+1
NO
≥19 nt IDENTITY?
YES
56
58
513
YES
57
514
GOI (+) mrk CSG^p
CONVERT (-) to (+)

FIGURE 5

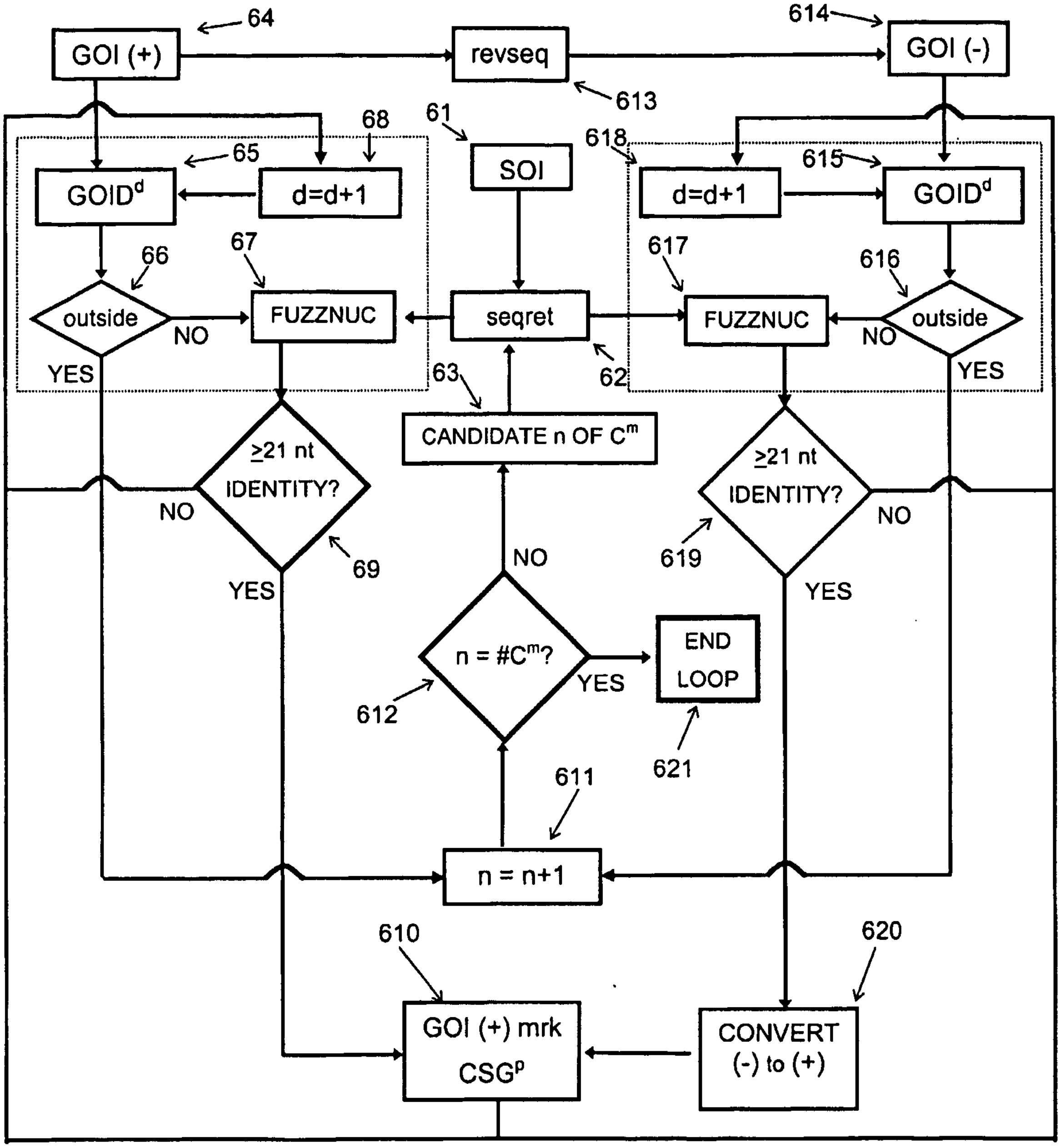

GOI (+)
64
revseq
613
614
GOI (-)
68
61
SOI
618
65
GOID^d
d=d+1
d=d+1
615
GOID^d
66
67
617
616
outside
NO
FUZZNUC
seqret
FUZZNUC
NO
outside
YES
63
62
YES
CANDIDATE n OF C^m
≥21 nt
IDENTITY?
NO
69
YES
NO
≥21 nt
IDENTITY?
619
NO
YES
n = #C^m?
YES
END
LOOP
612
621
611
n = n+1
610
620
GOI (+) mrk
CSG^p
CONVERT
(-) to (+)

FIGURE 6

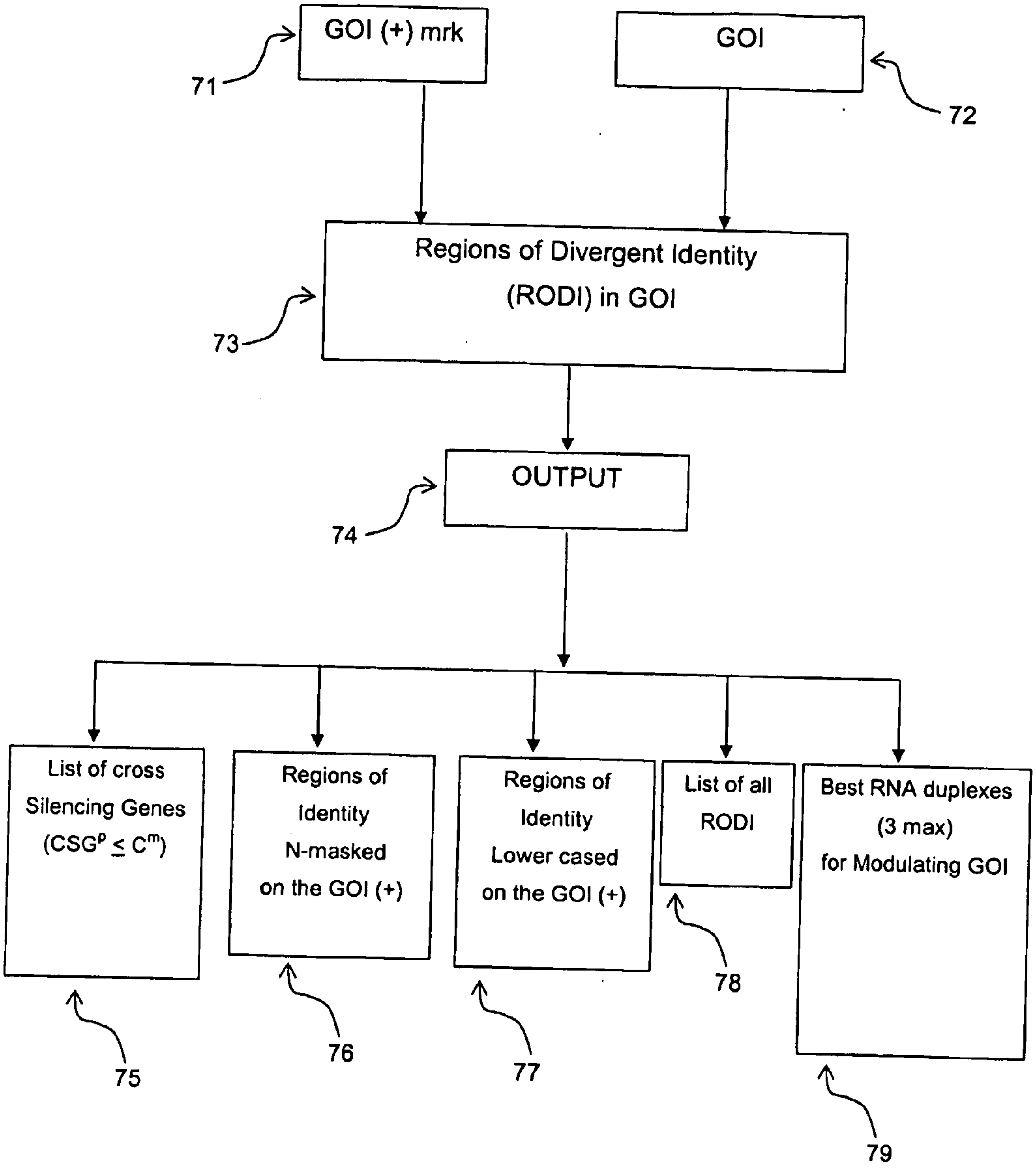

GOI (+) mrk
71
GOI
72
Regions of Divergent Identity (RODI) in GOI
73
OUTPUT
74
List of cross Silencing Genes ($CSG^p \leq C^m$)
75
Regions of Identity N-masked on the GOI (+)
76
Regions of Identity Lower cased on the GOI (+)
77
List of all RODI
78
Best RNA duplexes (3 max) for Modulating GOI
79

FIGURE 7

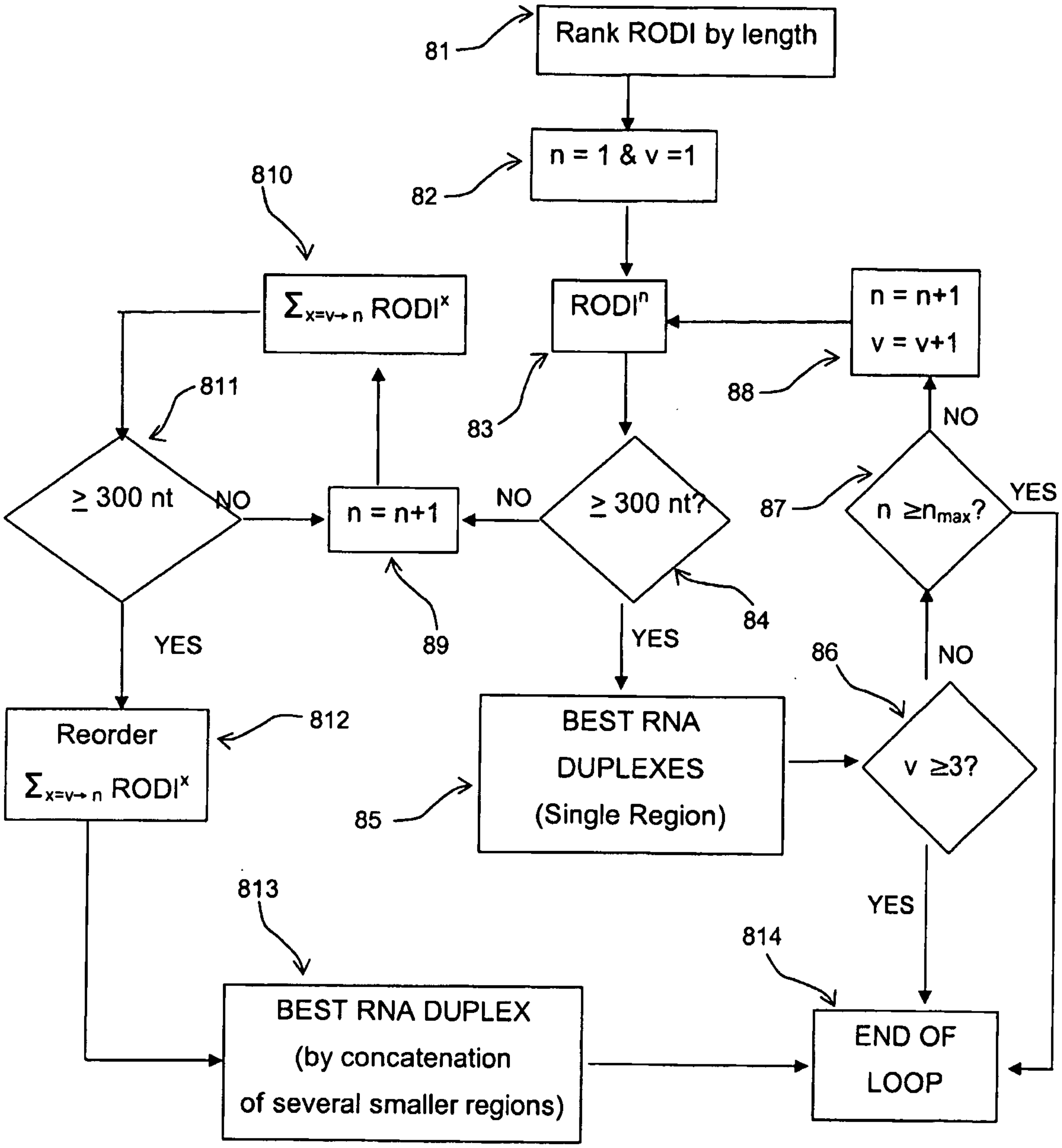

81
Rank RODI by length
82
n = 1 & v =1
810
$\Sigma_{x=v\rightarrow n}$ RODI^x
RODI^n
n = n+1
v = v+1
811
83
88
NO
≥ 300 nt
NO
n = n+1
NO
≥ 300 nt?
87
n ≥n$_{max}$?
YES
YES
89
YES
84
86
NO
Reorder
$\Sigma_{x=v\rightarrow n}$ RODI^x
812
BEST RNA DUPLEXES (Single Region)
85
v ≥3?
813
YES
814
BEST RNA DUPLEX (by concatenation of several smaller regions)
END OF LOOP

FIGURE 8

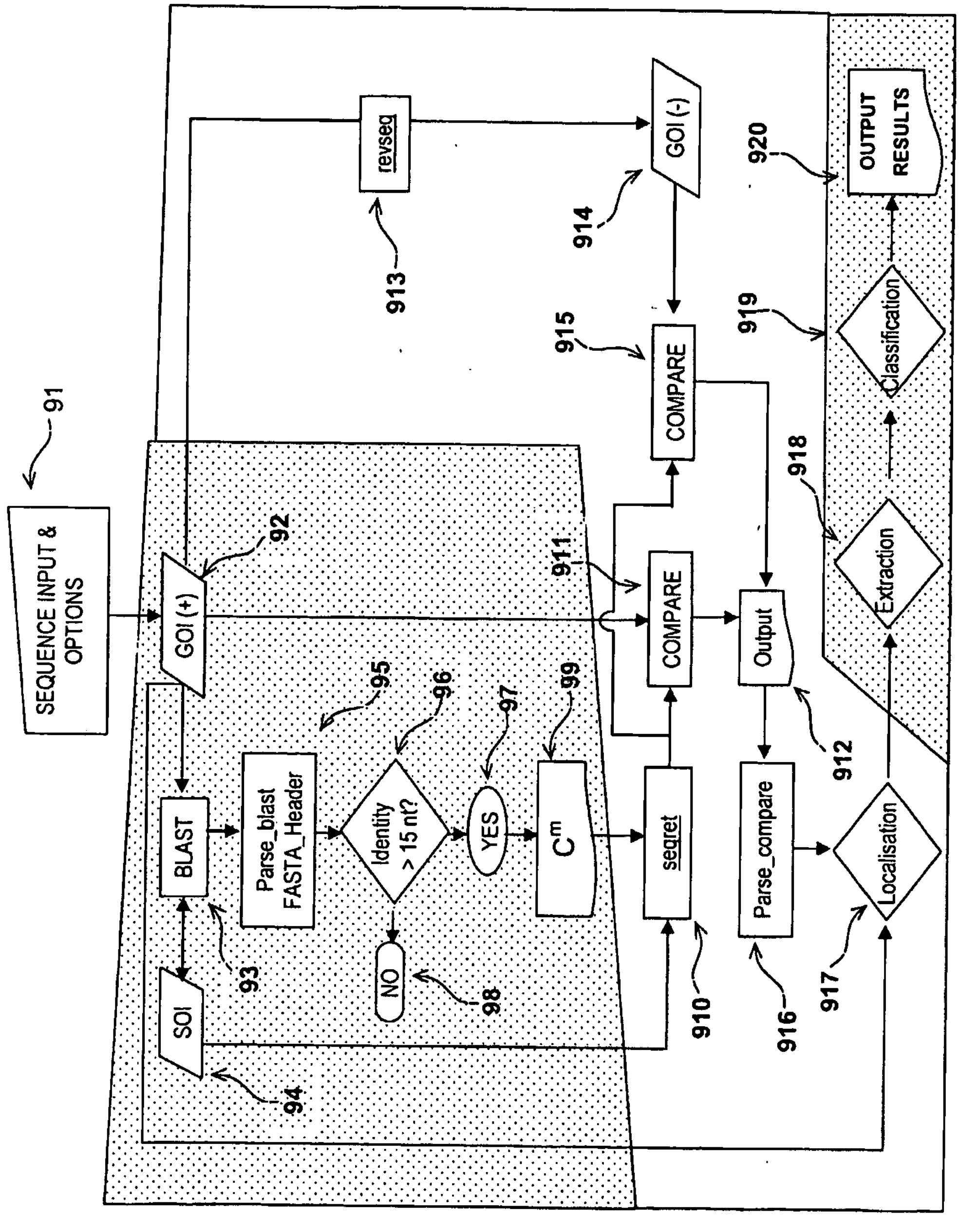

SEQUENCE INPUT & OPTIONS
91
GOI (+)
92
BLAST
93
SOI
94
Parse_blast FASTA_Header
95
Identity > 15 nt?
96
YES
97
NO
98
C^m
99
segret
910
COMPARE
911
Output
912
revseq
913
GOI (-)
914
COMPARE
915
Parse_compare
916
Localisation
917
Extraction
918
Classification
919
OUTPUT RESULTS
920

FIGURE 9

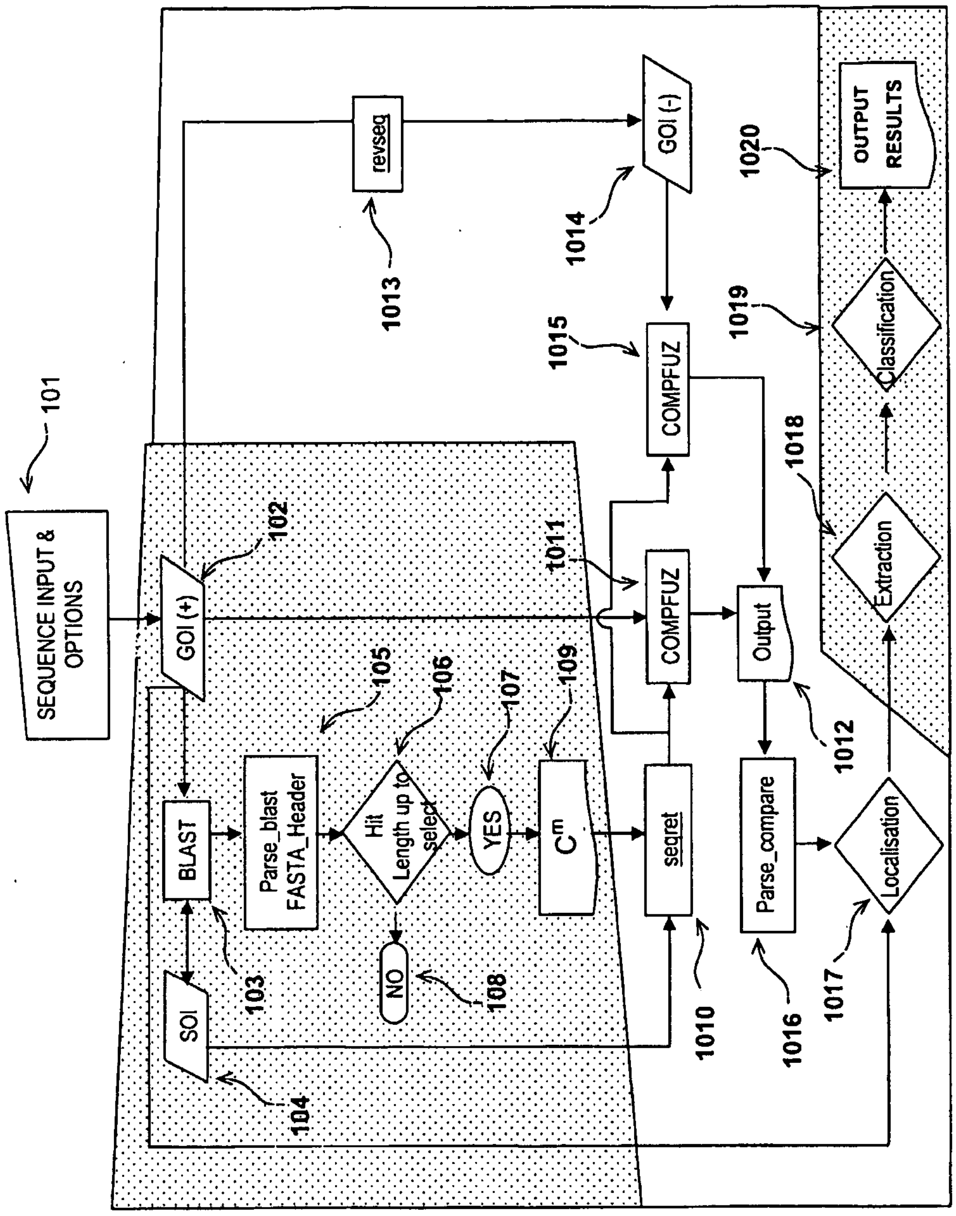

101
SEQUENCE INPUT & OPTIONS
102
GOI (+)
103
BLAST
104
SOI
105
Parse_blast FASTA_Header
106
Hit Length up to select
107
YES
108
NO
109
C^m
1010
seqret
1011
COMPFUZ
1012
Output
1013
revseq
1014
GOI (-)
1015
COMPFUZ
1016
Parse_compare
1017
Localisation
1018
Extraction
1019
Classification
1020
OUTPUT RESULTS

FIGURE 10

FGOI:

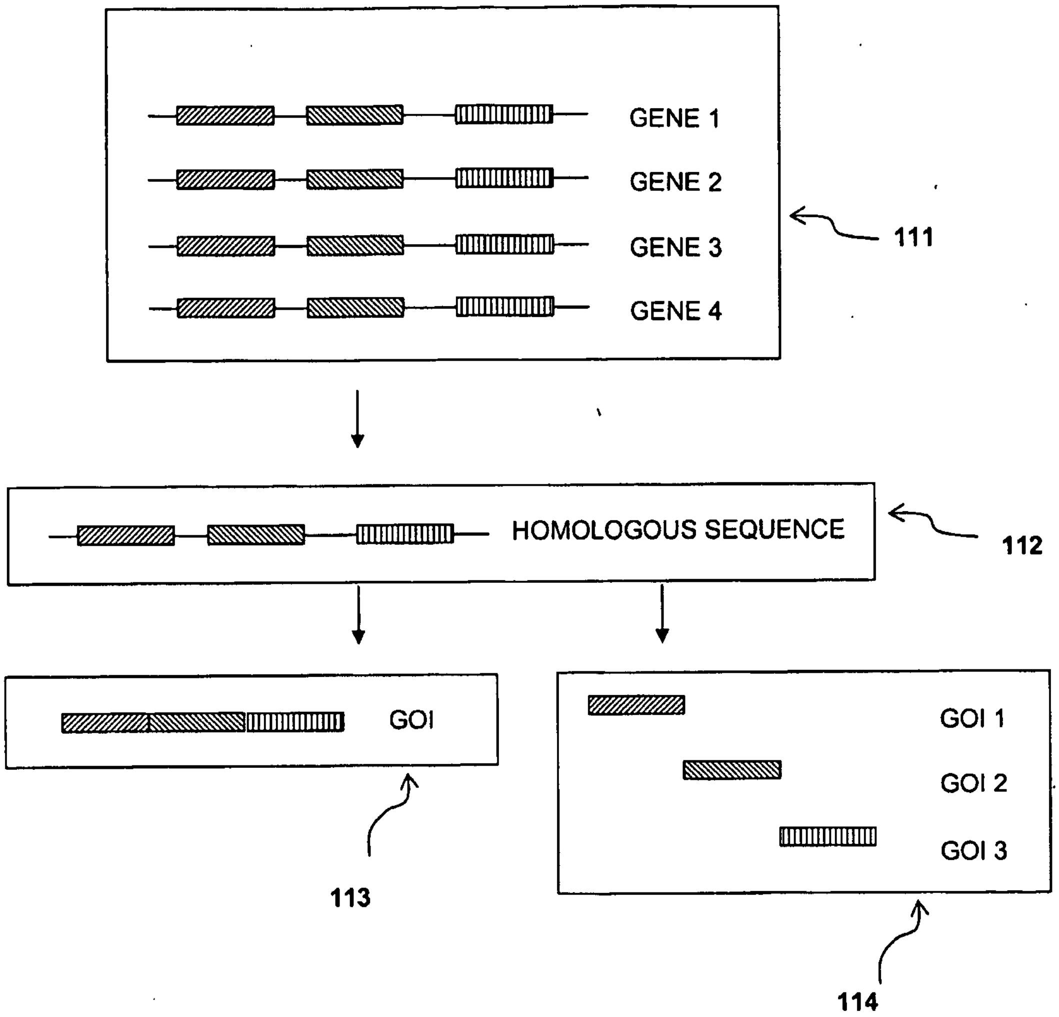


FIGURE 11

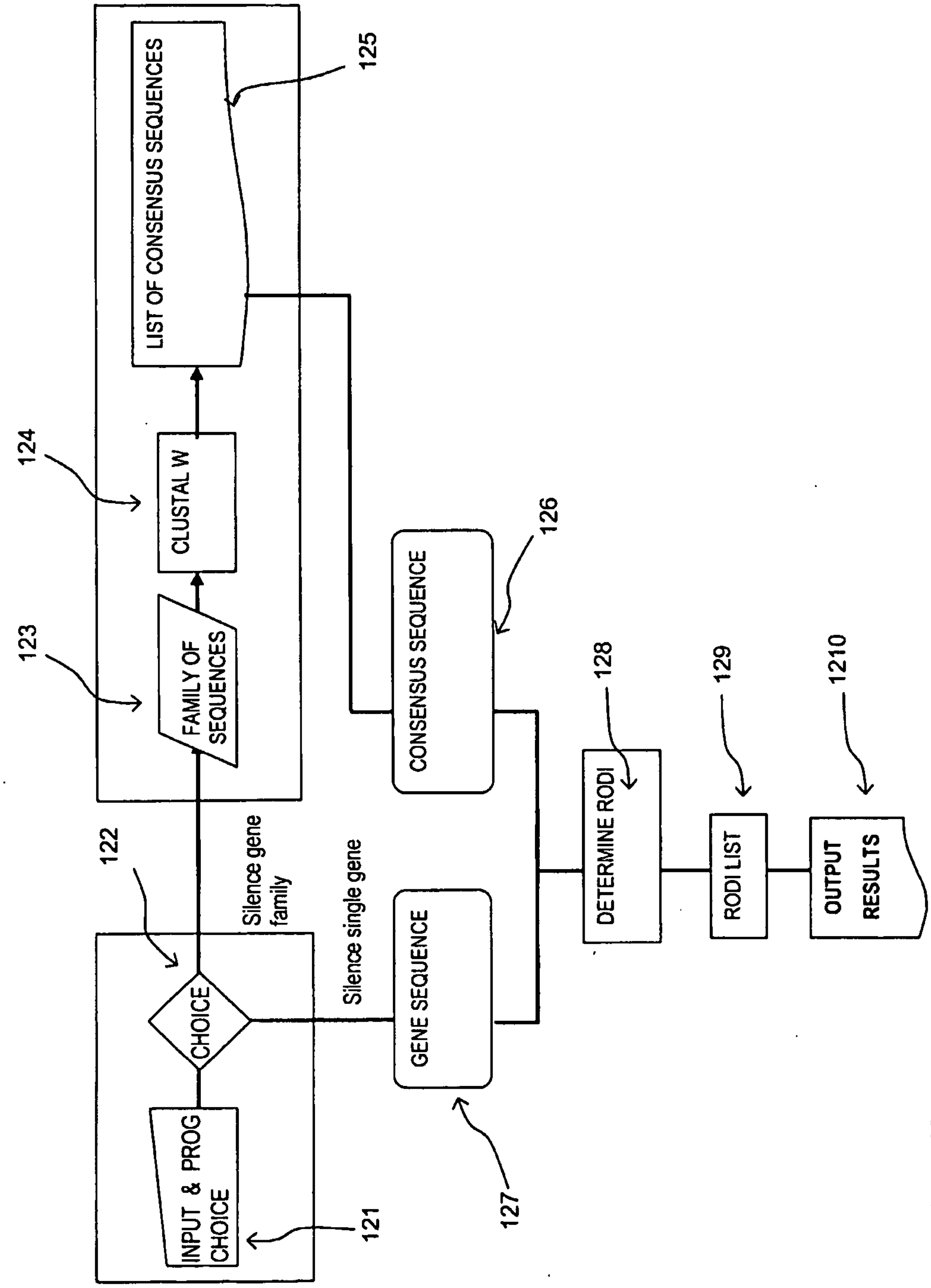

INPUT & PROG CHOICE
121
CHOICE
122
Silence gene family
Silence single gene
FAMILY OF SEQUENCES
123
CLUSTAL W
124
LIST OF CONSENSUS SEQUENCES
125
CONSENSUS SEQUENCE
126
GENE SEQUENCE
127
DETERMINE RODI
128
RODI LIST
129
OUTPUT RESULTS
1210

FIGURE 12

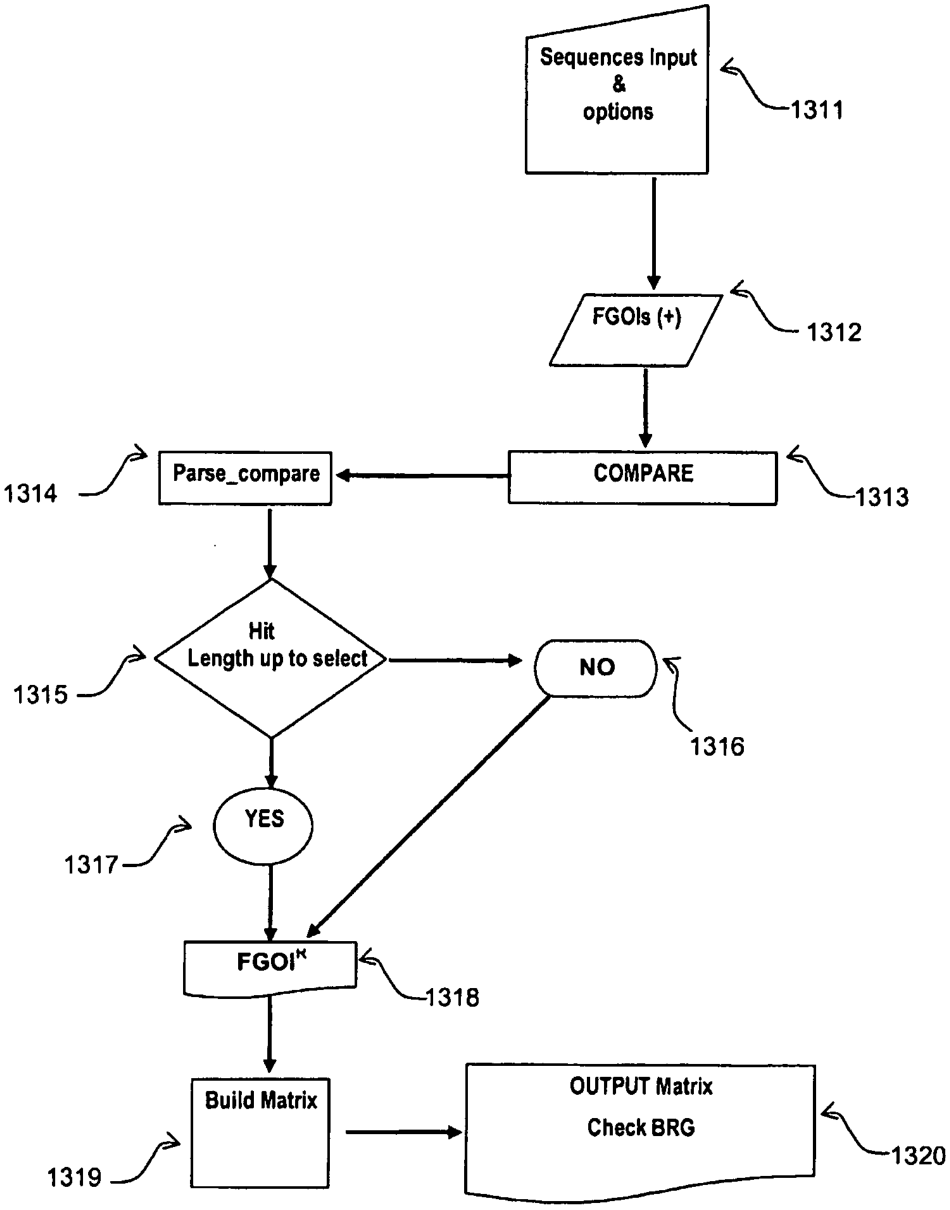

Sequences Input
&
options
1311
FGOIs (+)
1312
COMPARE
1313
Parse_compare
1314
Hit
Length up to select
1315
NO
1316
YES
1317
FGOI^K
1318
Build Matrix
1319
OUTPUT Matrix
Check BRG
1320

FIGURE 13A

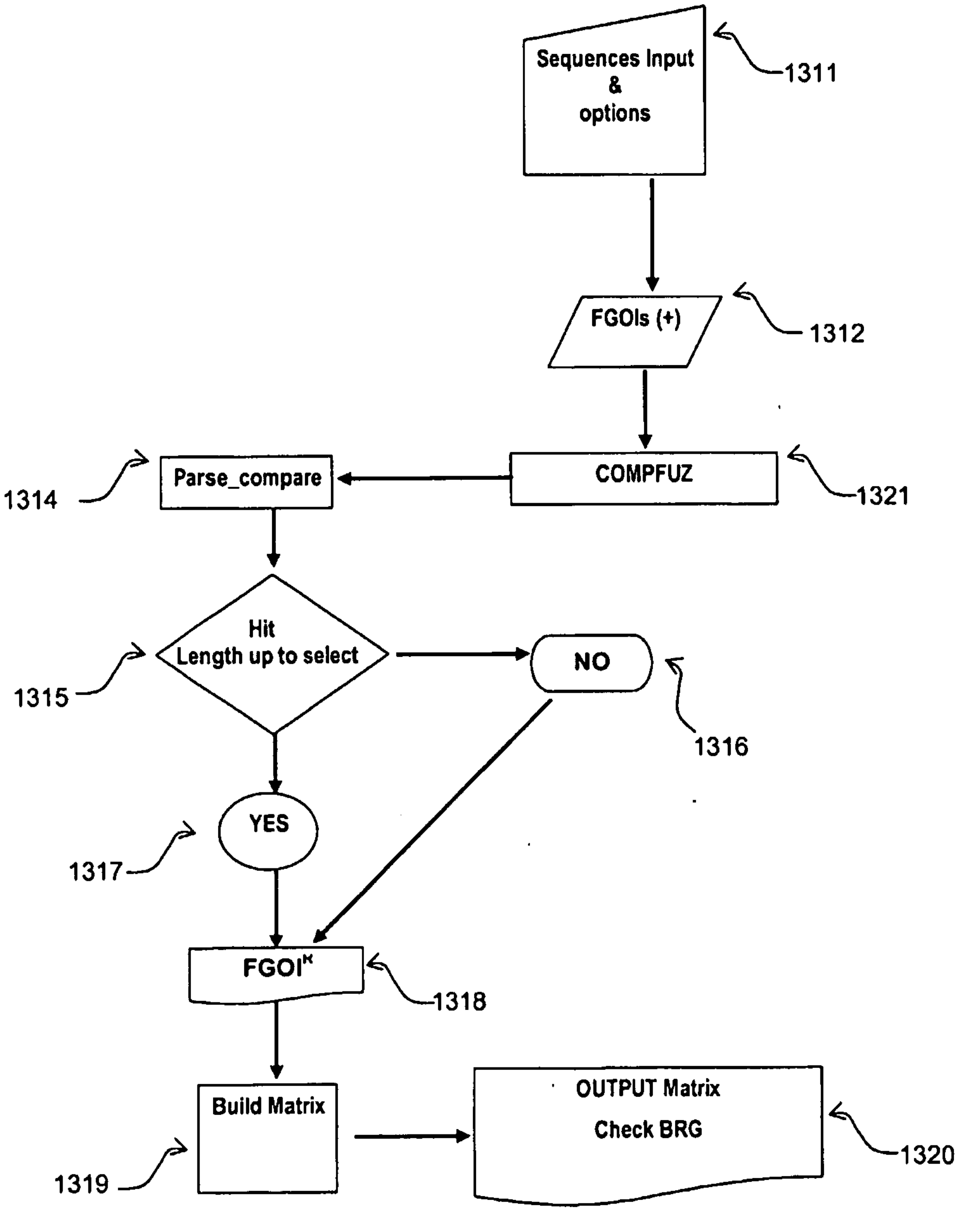

Sequences Input
&
options
1311
FGOIs (+)
1312
COMPFUZ
1321
Parse_compare
1314
Hit
Length up to select
1315
NO
1316
YES
1317
FGOI^K
1318
Build Matrix
1319
OUTPUT Matrix
Check BRG
1320

FIGURE 13B

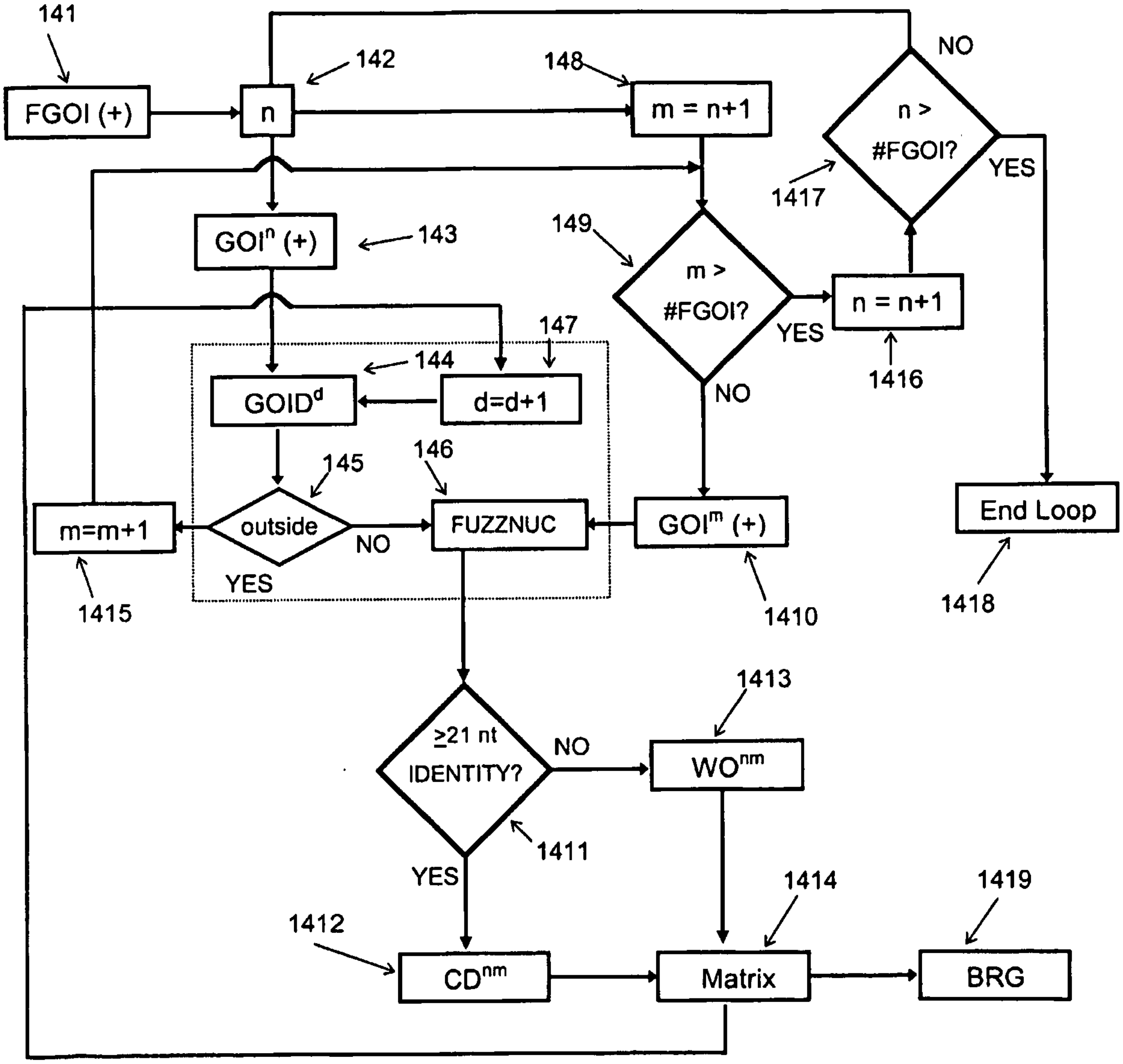
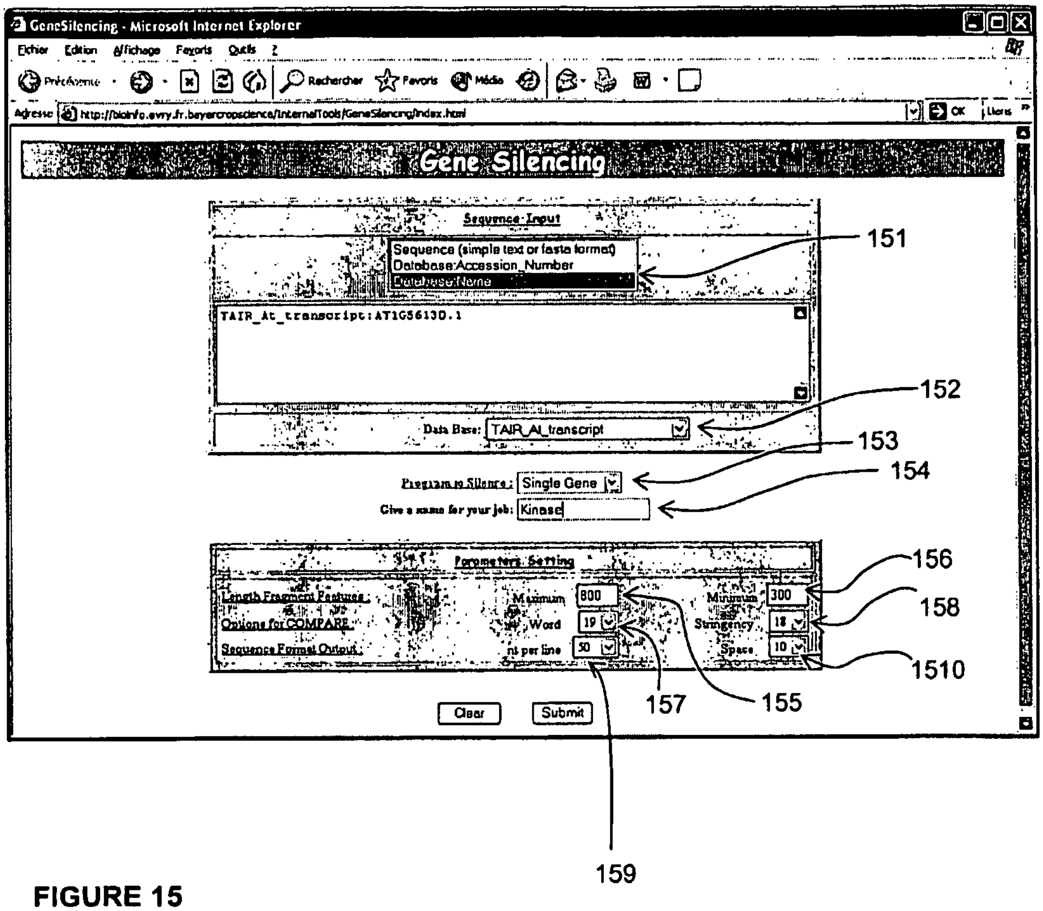

141
FGOI (+)
142
n
148
m = n+1
NO
n >
#FGOI?
YES
1417
143
GOI^n (+)
149
m >
#FGOI?
YES
n = n+1
1416
NO
147
144
GOID^d
d=d+1
145
146
m=m+1
outside
NO
FUZZNUC
GOI^m (+)
End Loop
YES
1415
1410
1418
1413
≥21 nt
IDENTITY?
NO
WO^{nm}
1411
YES
1412
1414
1419
CD^{nm}
Matrix
BRG

FIGURE 14



http://bioinfo.evry.fr.bayercropscience/InternalTools/GeneSilencing/index.html

Gene Silencing

Sequence Input

Sequence (simple text or fasta format)
Database:Accession_Number
Database:Name

TAIR_At_transcript:AT1G56130.1

Data Base: TAIR_At_transcript

Program to Silence: Single Gene

Give a name for your job: Kinase

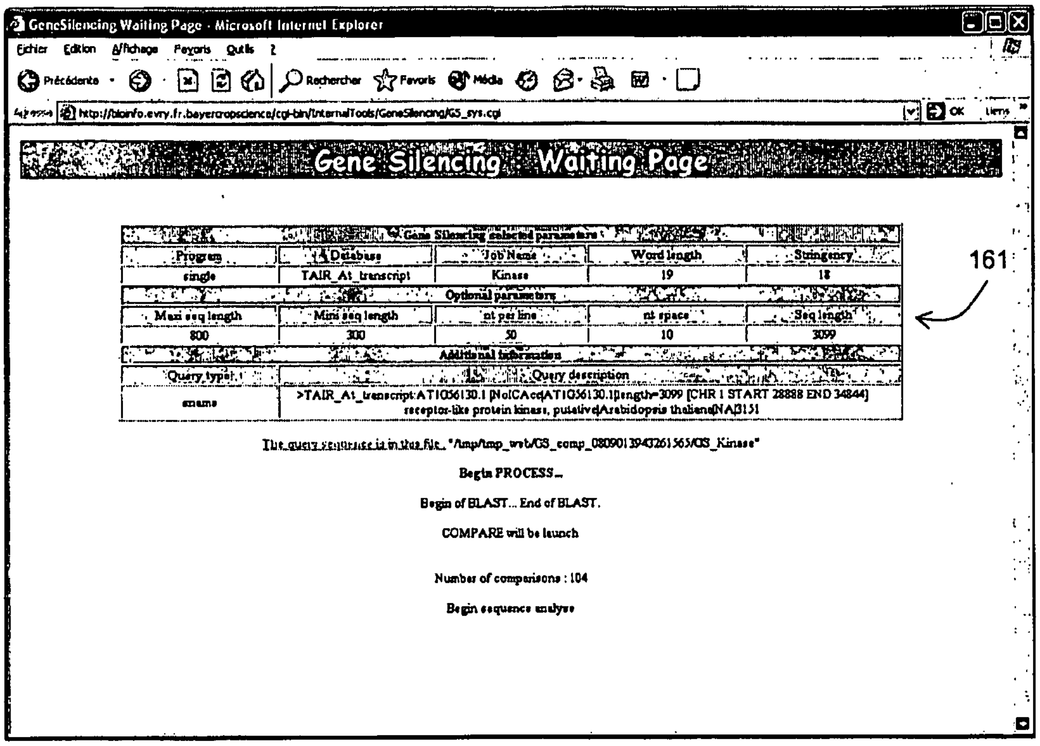
Parameters Setting

Length Fragment Features: Maximum 800 Minimum 300

Options for COMPARE: Word 19 Stringency 18

Sequence Format Output: nt per line 50 Space 10

FIGURE 15



GeneSilencing Waiting Page - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse http://bioinfo.evry.fr.bayercropscience/cgi-bin/InternalTools/GeneSilencing/GS_sys.cgi
OK
Liens
161

Gene Silencing : Waiting Page

Gene Silencing selected parameters				
Program	Database	Job Name	Word length	Stringency
single	TAIR_At_transcript	Kinase	19	18
Optional parameters				
Maxi seq length	Mini seq length	nt per line	nt space	Seq length
800	300	50	10	3099
Additional information				
Query type	Query description			
ename	>TAIR_At_transcript:AT1G56130.1 [NolCAcc[AT1G56130.1]length=3099 [CHR 1 START 28888 END 34844] receptor-like protein kinase, putative[Arabidopsis thaliana[NA]3151			

The query sequence is in this file.. "/tmp/tmp_web/GS_comp_080901394326156S/GS_Kinase"

Begin PROCESS...

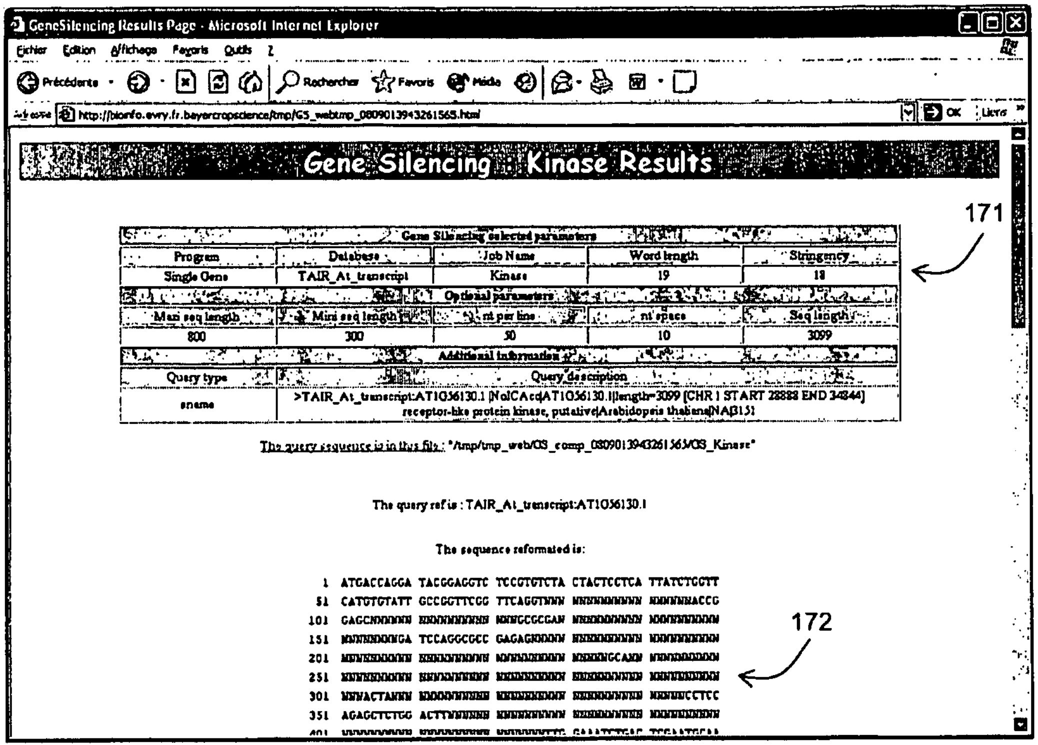
Begin of BLAST... End of BLAST.

COMPARE will be launch

Number of comparisons : 104

Begin sequence analyse

FIGURE 16



GeneSilencing Results Page - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse http://bioinfo.evry.fr.bayercropscience/tmp/GS_webtmp_0809013943261565.html
OK
Liens
171
172

Gene Silencing : Kinase Results

Gene Silencing selected parameters				
Program	Database	Job Name	Word length	Stringency
Single Gene	TAIR_At_transcript	Kinase	19	18
Optional parameters				
Max seq length	Mini seq length	nt per line	nt space	Seq length
800	300	50	10	3099
Additional information				
Query type	Query description			
sname	>TAIR_At_transcript:AT1G56130.1 [NolCAcd AT1G56130.1][length=3099 [CHR 1 START 28888 END 34844] receptor-like protein kinase, putative|Arabidopsis thaliana|NA|3151			

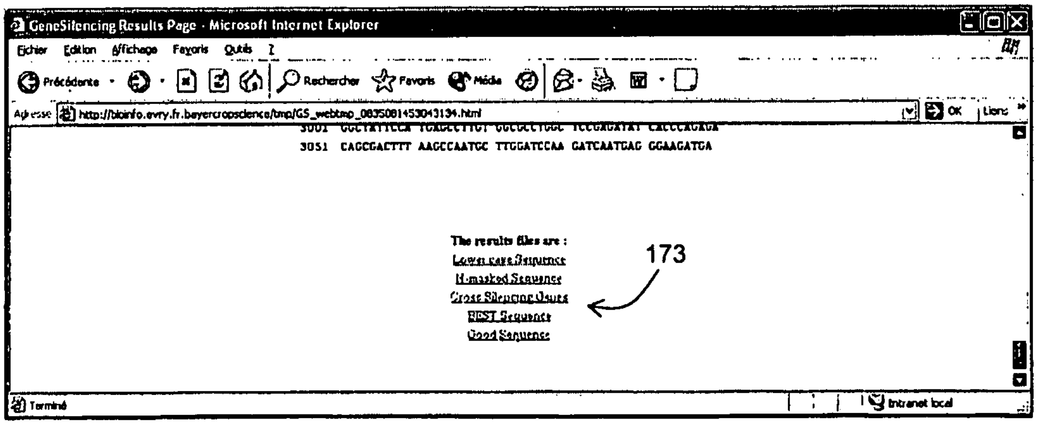
The query sequence is in this file : "/tmp/tmp_web/GS_comp_0809013943261565/GS_Kinase"

The query ref is : TAIR_At_transcript:AT1G56130.1

The sequence reformated is:

```
  1  ATGACCAGGA TACGGAGGTC TCCGTGTCTA CTACTCCTCA TTATCTGCTT
 51  CATGTGTATT GCCGGTTCGG TTCAGGTNNN NNNNNNNNNN NNNNNNACCG
101  GAGCNNNNNN NNNNNNNNNN NNNGCGCCAN NNNNNNNNNN NNNNNNNNNN
151  NNNNNNNNGA TCCAGGCGCC GAGAGNNNNN NNNNNNNNNN NNNNNNNNNN
201  NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNGCANN NNNNNNNNNN
251  NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN
301  NNNACTANNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNCCTCC
351  AGAGCTCTGG ACTTNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN
```

FIGURE 17-1



GeneSilencing Results Page - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse http://bioinfo.evry.fr.bayercropscience/tmp/GS_webtmp_0835081453043134.html
OK
Liens
Terminé
Intranet local
173

```
3051  CAGCGACTTT AAGCCAATGC TTGGATCCAA GATCAATGAG GGAAGATGA
```

The results files are :

Lower case Sequence

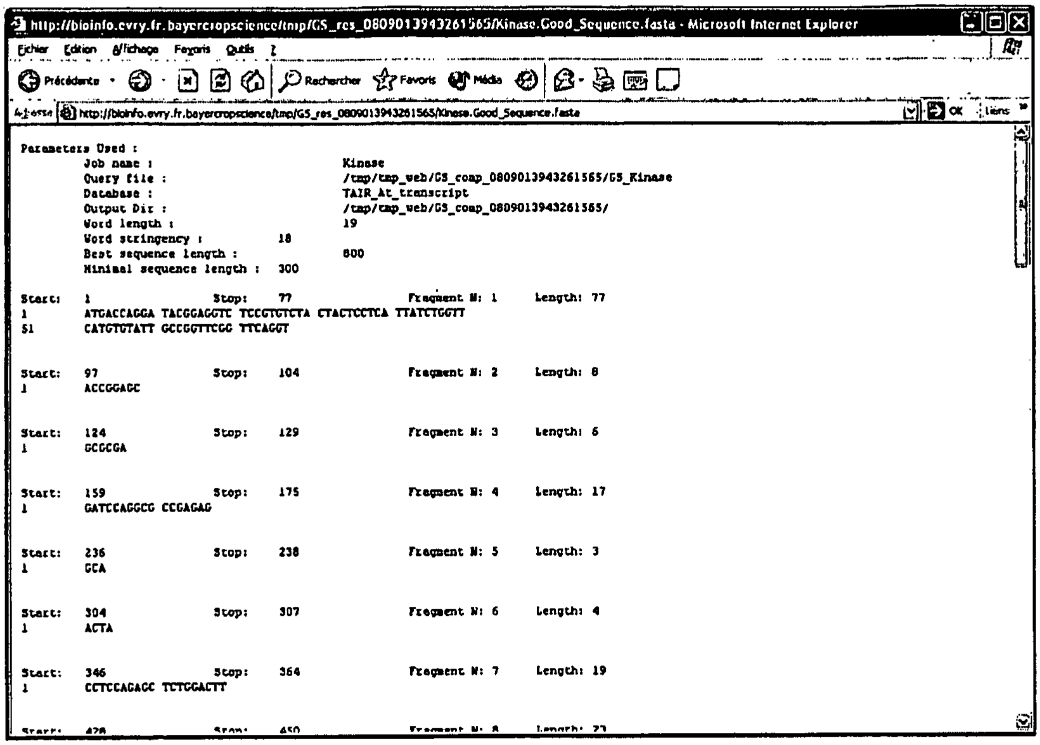
N-masked Sequence

Cross Silencing Genes

BEST Sequence

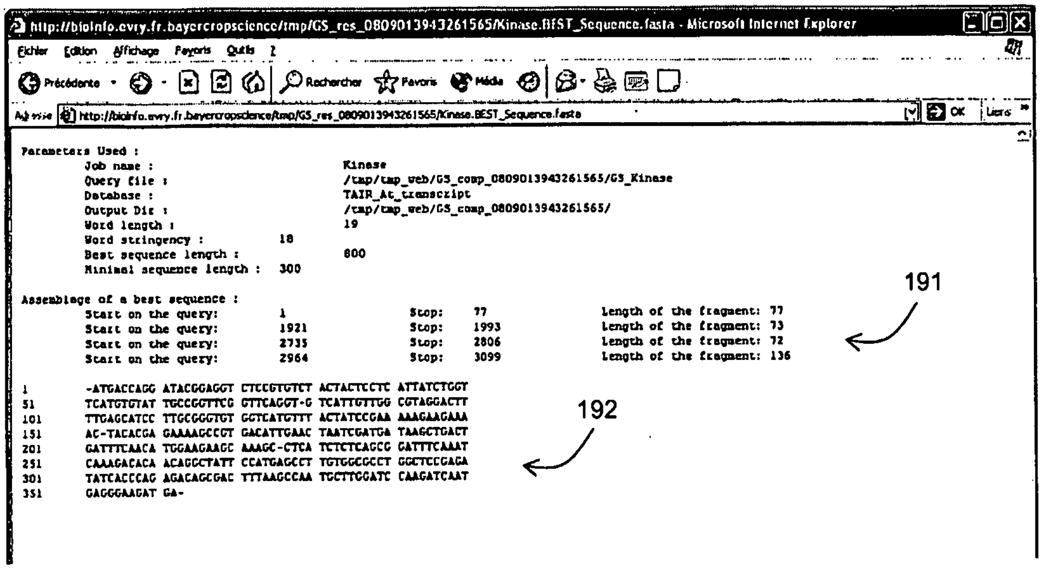
Good Sequence

FIGURE 17-2


http://bioinfo.evry.fr.bayercropscience/tmp/GS_res_0809013943261565/Kinase.Good_Sequence.fasta - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse
http://bioinfo.evry.fr.bayercropscience/tmp/GS_res_0809013943261565/Kinase.Good_Sequence.fasta
OK
Liens

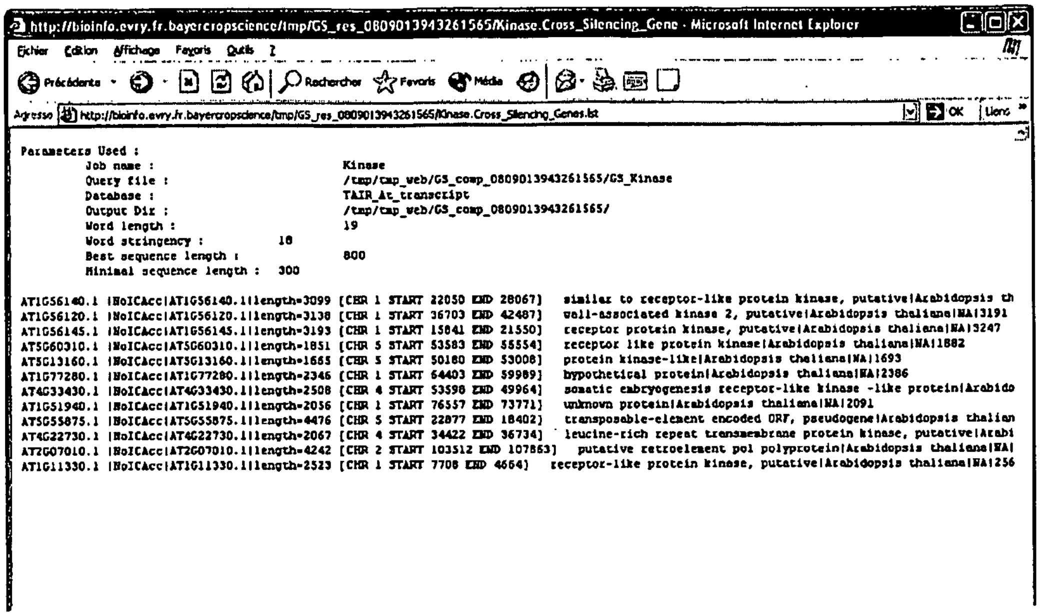
```
Parameters Used :
        Job name :                              Kinase
        Query file :                            /tmp/tmp_web/GS_comp_0809013943261565/GS_Kinase
        Database :                              TAIR_At_transcript
        Output Dir :                            /tmp/tmp_web/GS_comp_0809013943261565/
        Word length :                           19
        Word stringency :               18
        Best sequence length :                  800
        Minimal sequence length :       300

Start:  1               Stop:   77              Fragment N: 1   Length: 77
1       ATGACCAGGA TACGGAGGTC TCCGTGTCTA CTACTCCTCA TTATCTGGTT
51      CATGTGTATT GCCGGTTCGC TTCAGGT


Start:  97              Stop:   104             Fragment N: 2   Length: 8
1       ACCGGAGC


Start:  124             Stop:   129             Fragment N: 3   Length: 6
1       GCGCGA


Start:  159             Stop:   175             Fragment N: 4   Length: 17
1       GATCCAGGCG CCGAGAG


Start:  236             Stop:   238             Fragment N: 5   Length: 3
1       GCA


Start:  304             Stop:   307             Fragment N: 6   Length: 4
1       ACTA


Start:  346             Stop:   364             Fragment N: 7   Length: 19
1       CCTCCAGACC TCTGGACTT


Start:  428             Stop:   450             Fragment N: 8   Length: 23
```

FIGURE 18


http://bioinfo.evry.fr.bayercropscience/tmp/GS_res_0809013943261565/Kinase.BEST_Sequence.fasta - Microsoft Internet Explorer
Fichier
Edition
Affichage
Favoris
Outils
?
Précédente
Rechercher
Favoris
Média
Adresse
http://bioinfo.evry.fr.bayercropscience/tmp/GS_res_0809013943261565/Kinase.BEST_Sequence.fasta
OK
Liens
191
192

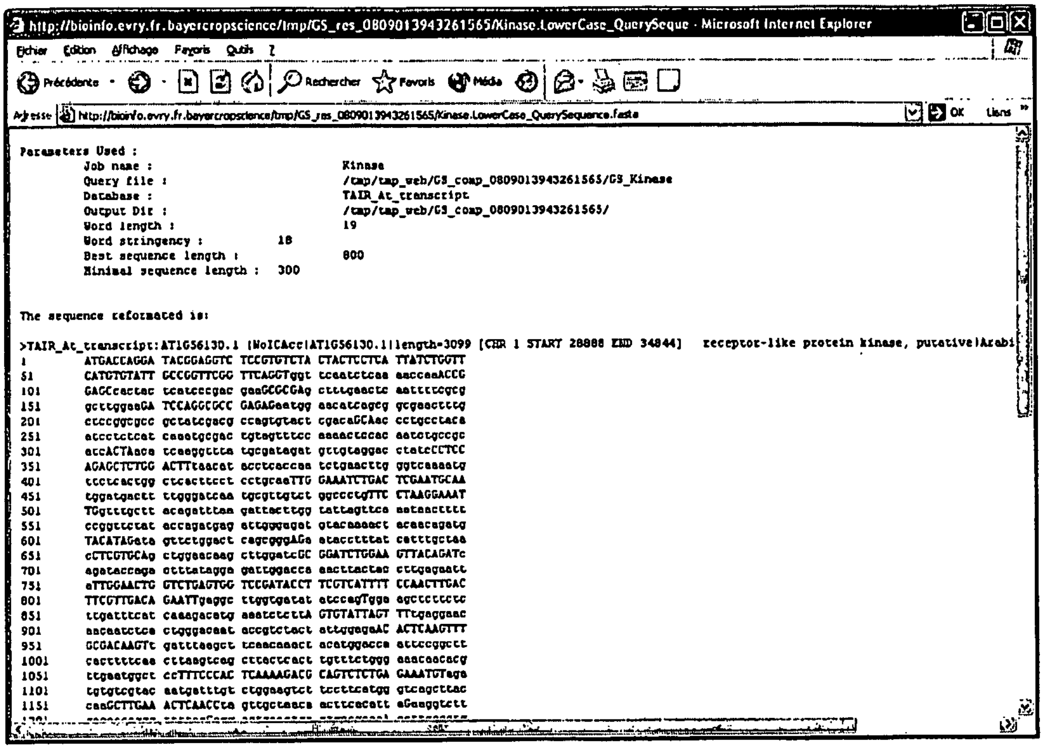
```
Parameters Used :
        Job name :                              Kinase
        Query file :                            /tmp/tmp_web/GS_comp_0809013943261565/GS_Kinase
        Database :                              TAIR_At_transcript
        Output Dir :                            /tmp/tmp_web/GS_comp_0809013943261565/
        Word length :                           19
        Word stringency :           18
        Best sequence length :                  800
        Minimal sequence length :   300

Assemblage of a best sequence :
        Start on the query:         1             Stop:    77          Length of the fragment: 77
        Start on the query:         1921          Stop:    1993        Length of the fragment: 73
        Start on the query:         2735          Stop:    2806        Length of the fragment: 72
        Start on the query:         2964          Stop:    3099        Length of the fragment: 136

1       -ATGACCAGG ATACGGAGGT CTCCGTGTCT ACTACTCCTC ATTATCTGGT
51      TCATGTGTAT TGCCGGTTCG GTTCAGGT-G TCATTGTTGG CGTAGGACTT
101     TTGAGCATCC TTGCGGGTGT GGTCATGTTT ACTATCCGAA AAAGAAGAAA
151     AC-TACACGA GAAAAGCCGT GACATTGAAC TAATCGATGA TAAGCTGACT
201     GATTTCAACA TGGAAGAAGC AAAGC-CTCA TCTCTCAGCG GATTTCAAAT
251     CAAAGACACA ACAGGCTATT CCATGAGCCT TGTGGCGCCT GGCTCCGAGA
301     TATCACCCAG AGACAGCGAC TTTAAGCCAA TGCTTGGATC CAAGATCAAT
351     GAGGGAAGAT GA-
```

FIGURE 19


http://bioinfo.evry.fr.bayercropscience/tmp/GS_res_0809013943261565/Kinase.Cross_Silencing_Gene - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse http://bioinfo.evry.fr.bayercropscience/tmp/GS_res_0809013943261565/Kinase.Cross_Silencing_Genes.lst
OK
Liens

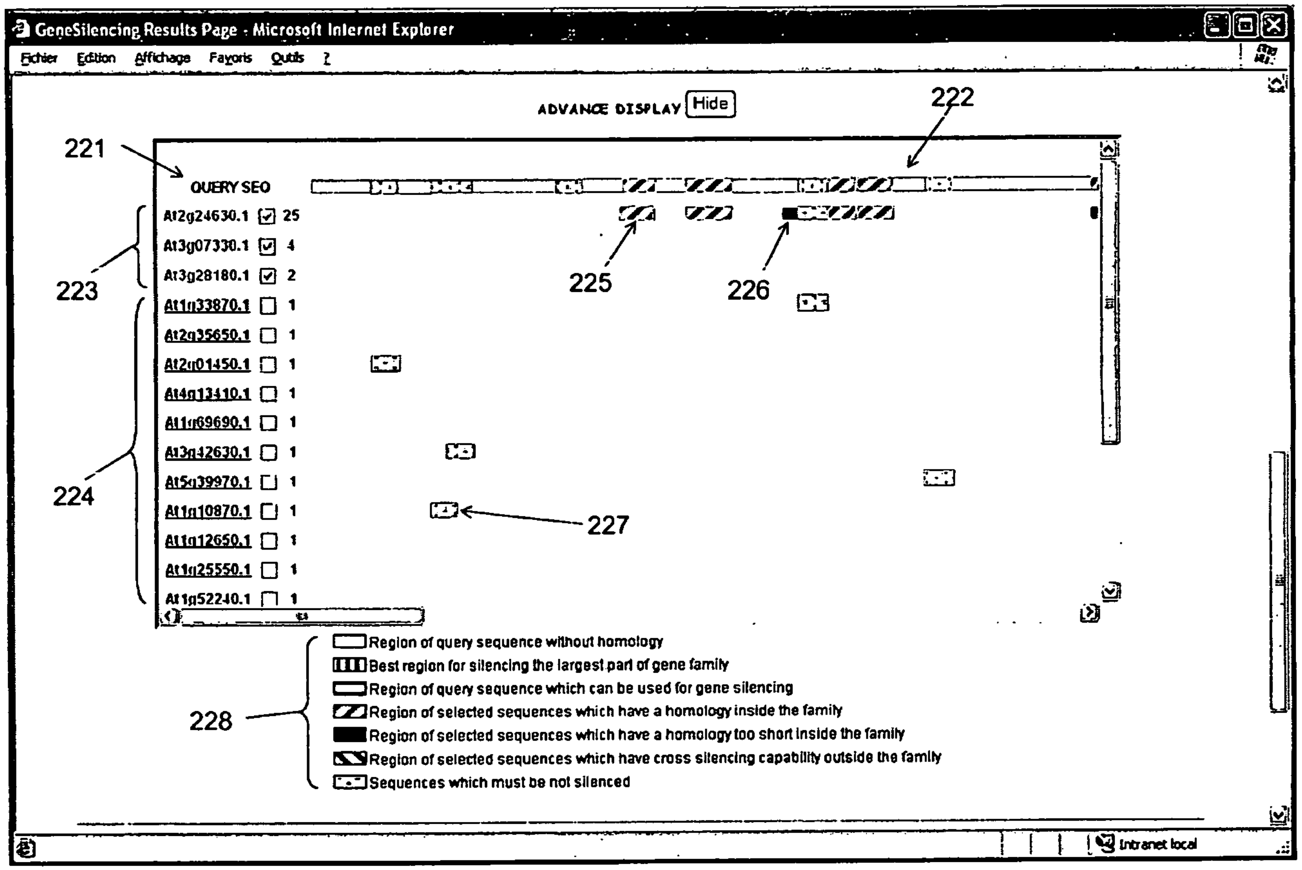
```
Parameters Used :
        Job name :                          Kinase
        Query file :                        /tmp/tmp_web/GS_comp_0809013943261565/GS_Kinase
        Database :                          TAIR_At_transcript
        Output Dir :                        /tmp/tmp_web/GS_comp_0809013943261565/
        Word length :                       19
        Word stringency :           18
        Best sequence length :              800
        Minimal sequence length :   300

AT1G56140.1 |NoICAcc|AT1G56140.1|length=3099 [CHR 1 START 22050 END 28067]   similar to receptor-like protein kinase, putative|Arabidopsis th
AT1G56120.1 |NoICAcc|AT1G56120.1|length=3138 [CHR 1 START 36703 END 42487]   wall-associated kinase 2, putative|Arabidopsis thaliana|NA|3191
AT1G56145.1 |NoICAcc|AT1G56145.1|length=3193 [CHR 1 START 15841 END 21550]   receptor protein kinase, putative|Arabidopsis thaliana|NA|3247
AT5G60310.1 |NoICAcc|AT5G60310.1|length=1851 [CHR 5 START 53583 END 55554]   receptor like protein kinase|Arabidopsis thaliana|NA|1882
AT5G13160.1 |NoICAcc|AT5G13160.1|length=1665 [CHR 5 START 50180 END 53008]   protein kinase-like|Arabidopsis thaliana|NA|1693
AT1G77280.1 |NoICAcc|AT1G77280.1|length=2346 [CHR 1 START 64403 END 59989]   hypothetical protein|Arabidopsis thaliana|NA|2386
AT4G33430.1 |NoICAcc|AT4G33430.1|length=2508 [CHR 4 START 53598 END 49964]   somatic embryogenesis receptor-like kinase -like protein|Arabido
AT1G51940.1 |NoICAcc|AT1G51940.1|length=2056 [CHR 1 START 76557 END 73771]   unknown protein|Arabidopsis thaliana|NA|2091
AT5G55875.1 |NoICAcc|AT5G55875.1|length=4476 [CHR 5 START 22877 END 18402]   transposable-element encoded ORF, pseudogene|Arabidopsis thalian
AT4G22730.1 |NoICAcc|AT4G22730.1|length=2067 [CHR 4 START 34422 END 36734]   leucine-rich repeat transmembrane protein kinase, putative|Arabi
AT2G07010.1 |NoICAcc|AT2G07010.1|length=4242 [CHR 2 START 103512 END 107863]   putative retroelement pol polyprotein|Arabidopsis thaliana|NA|
AT1G11330.1 |NoICAcc|AT1G11330.1|length=2523 [CHR 1 START 7708 END 4664]   receptor-like protein kinase, putative|Arabidopsis thaliana|NA|256
```

FIGURE 20


http://bioinfo.evry.fr.bayercropscience/tmp/GS_res_0809013943261565/Kinase.LowerCase_QuerySeque - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse http://bioinfo.evry.fr.bayercropscience/tmp/GS_res_0809013943261565/Kinase.LowerCase_QuerySequence.fasta
OK
Liens

```
Parameters Used :
        Job name :                            Kinase
        Query file :                          /tmp/tmp_web/GS_comp_0809013943261565/GS_Kinase
        Database :                            TAIR_At_transcript
        Output Dir :                          /tmp/tmp_web/GS_comp_0809013943261565/
        Word length :                         19
        Word stringency :           18
        Best sequence length :                800
        Minimal sequence length :   300


The sequence reformated is:

>TAIR_At_transcript:AT1G56130.1 |NoICAcc|AT1G56130.1||length=3099 [CHR 1 START 28888 END 34844]   receptor-like protein kinase, putative|Arabi
1       ATGACCAGGA TACGGAGGTC TCCGTGTCTA CTACTCCTCA TTATCTGGTT
51      CATGTGTATT GCCGGTTCGG TTCAGGTggt tcaatctcaa aaccaaACCG
101     GAGCcactac tcatcccgac gaaGCGCGAg ctttgaactc aattttcgcg
151     gcttggaaGA TCCAGGCGCC GAGAGaatgg aacatcagcg gcgaactttg
201     ctccggcgcc gctatcgacg ccagtgtact cgacaGCAac cctgcctaca
251     atcctctcat caaatgcgac tgtagtttcc aaaactccac aatctgccgc
301     atcACTAaca tcaaggttta tgcgatagat gttgtaggac ctatcCCTCC
351     AGAGCTCTGG ACTTtaacat acctcaccaa tctgaacttg ggtcaaaatg
401     tcctcactgg ctcactttct cctgcaaTTG GAAATCTGAC TCGAATGCAA
451     tggatgactt ttgggatcaa tgcgttgtct ggccctgTTC CTAAGGAAAT
501     TGgtttgctt acagatttaa gattacttgg tattagttca aataactttt
551     ccggttctat accagatgag attgggagat gtacaaaact acaacagatg
601     TACATAGata gttctggact cagcgggAGa atacctttat catttgctaa
651     cCTCGTGCAg ctggaacaag cttggatcGC GGATCTGGAA GTTACAGATc
701     agataccaga atttatagga gattggacca aacttactac cttgagaatt
751     aTTGGAACTG GTCTGAGTGG TCCGATACCT TCGTCATTTT CCAACTTGAC
801     TTCGTTGACA GAATTgaggc ttggtgatat atccagTgga agctcttctc
851     ttgatttcat caaagacatg aaatctcttA GTGTATTAGT TTtgaggaac
901     aacaatctca ctgggacaat accgtctact attggagaAC ACTCAAGTTT
951     GCGACAAGTt gatttaagct tcaacaaact acatggacca attccggctt
1001    cactttttca cttaagtcag cttactcact tgtttctggg aaacaacacg
1051    ttgaatggct ccTTTCCCAC TCAAAAGACG CAGTCTCTGA GAAATGTaga
1101    tgtgtcgtac aatgatttgt ctggaagtct tccttcatgg gtcagcttac
1151    caaGCTTGAA ACTCAACCta gttgctaaca acttcacatt aGaaggtctt
```

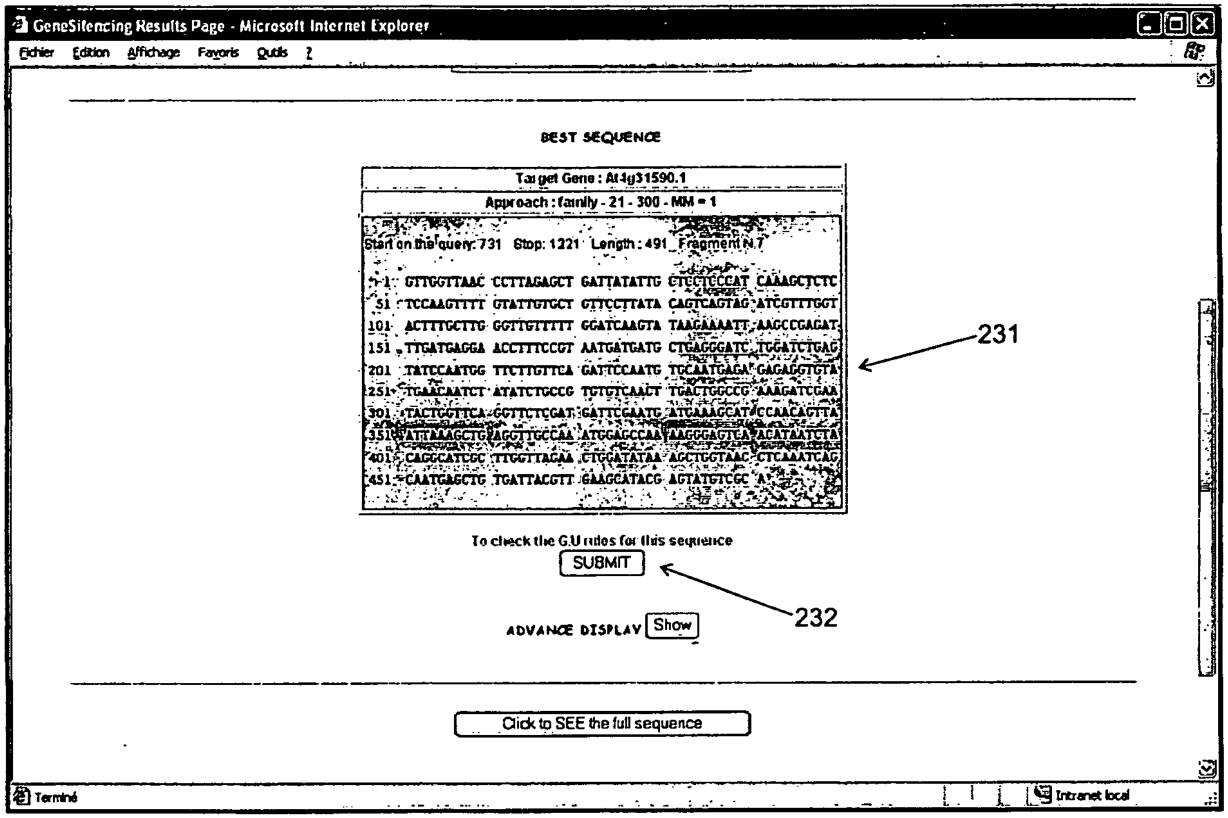
FIGURE 21



QUERY SEQ		
At2g24630.1	☑	25
At3g07330.1	☑	4
At3g28180.1	☑	2
At1g33870.1	□	1
At2g35650.1	□	1
At2g01450.1	□	1
At4g13410.1	□	1
At1g69690.1	□	1
At3g42630.1	□	1
At5g39970.1	□	1
At1g10870.1	□	1
At1g12650.1	□	1
At1g25550.1	□	1
At1g52240.1	□	1

□ Region of query sequence without homology
▥ Best region for silencing the largest part of gene family
□ Region of query sequence which can be used for gene silencing
▨ Region of selected sequences which have a homology inside the family
■ Region of selected sequences which have a homology too short inside the family
▧ Region of selected sequences which have cross silencing capability outside the family
⬚ Sequences which must be not silenced

FIGURE 22


GeneSilencing Results Page - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
231
232
Terminé
Intranet local

BEST SEQUENCE

Target Gene : At4g31590.1

Approach : family - 21 - 300 - MM = 1

```
Start on the query: 731   Stop: 1221   Length : 491   Fragment N.7

  1  GTTGGTTAAC CCTTAGAGCT GATTATATTG CTCCTCCCAT CAAAGCTCTC
 51  TCCAAGTTTT GTATTGTGCT GTTCCTTATA CAGTCAGTAG ATCGTTTGGT
101  ACTTTGCTTG GGTTGTTTTT GGATCAAGTA TAAGAAAATT AAGCCGAGAT
151  TTGATGAGGA ACCTTTCCGT AATGATGATG CTGAGGGATC TGGATCTGAG
201  TATCCAATGG TTCTTGTTCA GATTCCAATG TGCAATGAGA GAGAGGTGTA
251  TGAACAATCT ATATCTGCCG TGTGTCAACT TGACTGGCCG AAAGATCGAA
301  TACTGGTTCA GGTTCTCGAT GATTCGAATG ATGAAAGCAT CCAACAGTTA
351  ATTAAAGCTG AGGTTGCCAA ATGGAGCCAA AAGGGAGTCA ACATAATCTA
401  CAGGCATCGC TTGGTTAGAA CTGGATATAA AGCTGGTAAC CTCAAATCAG
451  CAATGAGCTG TGATTACGTT GAAGCATACG AGTATGTCGC A
```

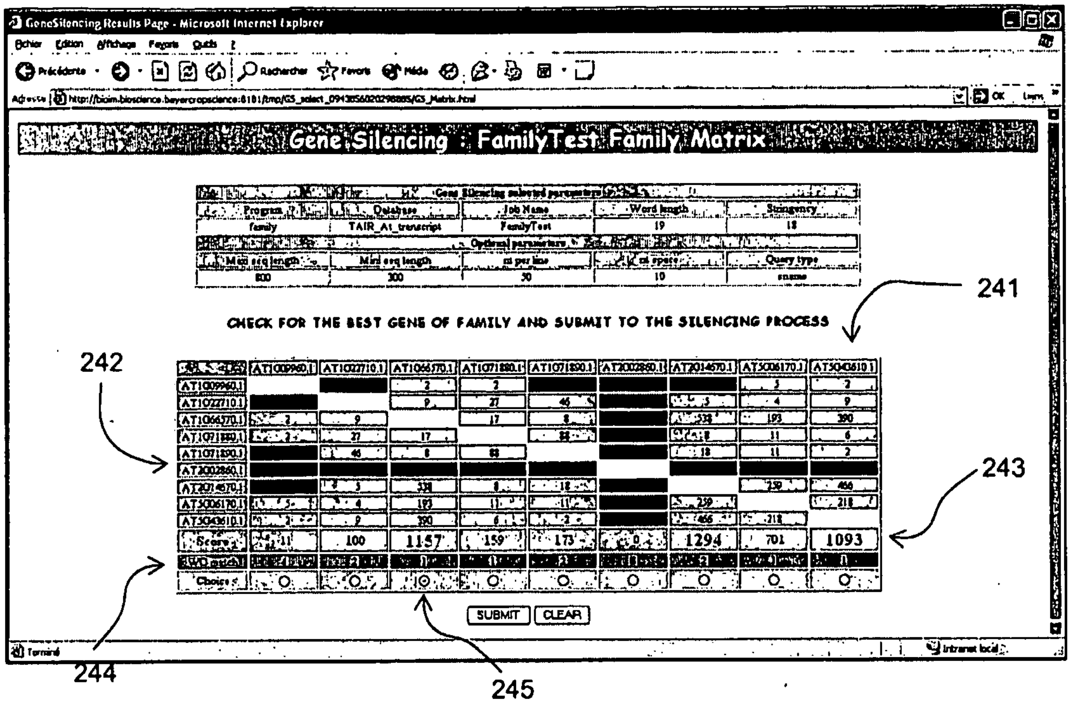
To check the G.U rules for this sequence

SUBMIT

ADVANCE DISPLAY Show

Click to SEE the full sequence

FIGURE 23


GeneSilencing Results Page - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse http://bioim.bioscience.bayercropscience:8181/tmp/GS_select_0943856020298885/GS_Matrix.html
OK
Liens
Terminé
Intranet local
241
242
243
244
245

Gene Silencing : FamilyTest Family Matrix

Gene Silencing selected parameters				
Program	Database	Job Name	Word length	Stringency
family	TAIR_At_transcript	FamilyTest	19	18
Optional parameters				
Maxi seq length	Mini seq length	nt per line	nt space	Query type
800	300	50	10	sname

CHECK FOR THE BEST GENE OF FAMILY AND SUBMIT TO THE SILENCING PROCESS

[illegible]	AT1009960.1	AT1022710.1	AT1066570.1	AT1071880.1	AT1071890.1	AT2002860.1	AT2014670.1	AT5006170.1	AT5040610.1
AT1009960.1		■	2	2	■	■	■	5	2
AT1022710.1	■		9	27	46	■	5	4	9
AT1066570.1	2	9		17	8	■	538	193	390
AT1071880.1	2	27	17		88	■	8	11	6
AT1071890.1	■	46	8	88		■	18	11	2
AT2002860.1	■	■	■	■	■		■	■	■
AT2014670.1	■	5	538	8	18	■		259	466
AT5006170.1	5	4	193	11	11	■	259		218
AT5040610.1	2	9	390	6	2	■	466	218	
Score	11	100	1157	159	173	0	1294	701	1093
SWO match	[illegible]	2	0	0	2	0	2	0	0
Choice	○	○	●	○	○	○	○	○	○

SUBMIT CLEAR

FIGURE 24